Advancing Ethical AI: Methods for fairness enhancement leveraging on causality and under privacy constraints

Rūta Binkytė

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAX145

ÉCOLE POLYTECHNIQUE

IP PARIS

Advancing Ethical AI

Methods for fairness enhancement leveraging on causality and under privacy constraints

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Mathématiques et informatique

Thèse présentée et soutenue à Palaiseau, le 18 décembre 2023, par

RŪTA BINKYTĖ-SADAUSKIENĖ

Composition du Jury :

Stefan HAAR Directeur de recherche Inria Saclay-Ile-de-Franc	Président
Mario FRITZ Directeur de recherche CISPA Helmholtz Center for Information Security	Rapporteur et examinateur
Pascal VAN HENTENRYCK Full professor H. Milton Stewart School of Industrial and Systems Engineering	Rapporteur non examinateur
Miguel COUCEIRO Professor Université de Lorraine	Examinateur
Sébastien GAMBS Professeur Université du Québec à Montréal Département d'informatique	Examinateur
Daniele GORLA Associate Professor Sapienza Università di Roma	Examinateur
Catuscia Palamidessi Directrice de recherche, Inria (Comète)	Directrice de thèse
Sami Zhioua Chercheur, Inria (Comète)	Co-directeur de thèse

Abstract

Ethical [1] Artificial intelligence (AI) is a set of practices and theories within the field of AI and machine learning (ML) that aim to align AI practices with moral values, addressing the potential impact on human lives. The ethical principles that technologies are expected to satisfy may vary between the organizations and countries that formulated them. There are multiple versions of possible ethical guidelines for ethical AI. Some of the common ethical AI principles include [1, 2]:

- fairness and non-discrimination,
- privacy,
- interpretability and explainability,
- safety and reliability.

Our work focuses on fairness, causality, and privacy in the context of machine learning and AI systems. We see causality as a tool that can soften trade-offs and foster synergies among diverse aspects of ethical AI. The general contribution of this thesis includes the holistic approach to Ethical AI, which aims to incorporate methods for fairness, privacy, and causality. The specific contributory elements are structured around chapters, each addressing specific aspects:

Chapter 7: BaBE - Enhancing Fairness via Estimation of Latent Explaining Variables proposes a novel approach using Expectation-Maximization to estimate the conditional distribution of a latent explaining variable, laying the foundation for data pre-processing to improve fairness, accuracy, and explainability.

Chapter 8: Underrepresentation and Sampling Bias in Machine Learning systematically analyzes the impact of sample size and underrepresentation on discrimination in algorithmic decisions, linking the analysis to bias mitigation techniques.

Chapter 9: Causal Discovery under Local Privacy compares the performance of different locally private mechanisms in the context of causal discovery tasks, emphasizing the advantages of using geometric obfuscation methods.

Chapter 10: On the Need and Applicability of Causality for Fair Machine Learning emphasizes the necessity of incorporating causality into fair AI, connecting causality in fair AI with European AI legislation and discussing practical requirements.

Chapter 11: Dissecting Causal Biases explores causal biases and develops closed-form expressions for various sources, including confounding, colliding, measurement, and introduces interaction bias.

[1] Other terms include 'Trustworthy', 'Reliable', 'Human Centric'

Chapter 12: Gender and Sex Bias in COVID-19 Data provides a comprehensive review of the literature on gender and sex bias in COVID-19 data, using causal graphs for analysis and emphasizing the importance of explainability and causality.

The thesis contributes to the field by offering insights and novel approaches for achieving better trade-offs between fairness, accuracy, privacy, and explainability in AI systems. It suggests future research directions and advocates for a holistic approach to ethical AI development.

Résumé

Intelligence Artificielle Éthique (IA) regroupe un ensemble de pratiques et de théories dans le domaine de l'IA et de l'apprentissage automatique visant à aligner les pratiques de l'IA sur des valeurs morales, en abordant l'impact potentiel sur la vie humaine. Les principes éthiques que les technologies sont censées satisfaire peuvent varier selon les organisations et les pays qui les ont formulés. Il existe plusieurs versions de directives éthiques possibles pour une IA éthique. Certains des principes éthiques communs de l'IA incluent:

- équité et non-discrimination,
- interprétabilité et explicabilité,
- sécurité et fiabilité.

Notre travail se concentre sur l'équité, la causalité et la vie privée dans le contexte de l'apprentissage automatique et des systèmes d'IA. Nous considérons la causalité comme un outil capable d'atténuer les compromis et de favoriser des synergies entre les divers aspects de l'IA éthique. La contribution générale de cette thèse inclut une approche holistique de l'IA éthique, qui vise à incorporer des méthodes pour l'équité, la vie privée et la causalité. Les éléments spécifiques de contribution sont structurés autour des chapitres, chacun abordant des aspects spécifiques:

Chapitre 7: BaBE - Amélioration de l'équité via l'estimation des variables explicatives latentes propose une nouvelle approche utilisant l'espérance-maximisation pour estimer la distribution conditionnelle d'une variable explicative latente, jetant les bases du prétraitement des données pour améliorer l'équité, la précision et l'explicabilité.

Chapitre 8: Sous-représentation et biais d'échantillonnage en apprentissage automatique analyse systématiquement l'impact de la taille de l'échantillon et de la sous-représentation d'un groupe sur la discrimination dans les décisions algorithmiques, liant l'analyse aux techniques d'atténuation des biais. - Chapitre 9: Découverte causale sous la confidentialité locale compare les performances de différents mécanismes localement privés dans le contexte des tâches de découverte causale, mettant en avant les avantages des méthodes d'obfuscation géométrique.

Chapitre 10: Sur la nécessité et l'applicabilité de la causalité pour un apprentissage automatique équitable souligne la nécessité d'incorporer la causalité dans une IA équitable, reliant la causalité dans l'IA équitable à la législation européenne sur l'IA et discutant des exigences pratiques.

Chapitre 11: Dissection des biais causaux explore les biais causaux et développe des expressions mathématiques pour diverses sources, notamment la confusion, la collision, la mesure, et introduit le biais d'interaction.

Chapitre 12: Biais de genre et de sexe dans les données COVID-19 offre une revue complète de la littérature sur les biais de genre et de sexe dans les données COVID-19, utilisant des graphiques

causaux pour l'analyse et soulignant l'importance de l'explicabilité et de la causalité.

La thèse contribue au domaine en offrant des perspectives et des approches novatrices pour obtenir de meilleurs compromis entre l'équité, la précision, la vie privée et l'explicabilité dans les systèmes d'IA. Elle suggère des orientations de recherche future et plaide en faveur d'une approche holistique du développement de l'IA éthique axée sur les données.

Acknowledgements

I would like to express deepest gratitude to the many people who have made my PhD journey not only possible, but also an enjoyable experience.

To my supervisor **Catuscia Palamidessi**, who believed in me and my abilities more than I did myself. I deeply admire her scientific sharpness, creativity, sense of humor, and kindness. She has an almost supernatural ability to endure the sleepless nights before the submissions. She is also an incredibly talented manager, able to create an atmosphere of inclusion, support, and ease in the team. Last, but not the least, she makes the best lasagna in the world!

To my supervisor **Sami Zhioua**, an incredibly talented teacher and a dear friend. Sami has taught me how to navigate scientific complexity and was always able to guide without criticizing or discouraging me. Besides that, I admire him for being an accomplished athlete and an avid reader, who is not afraid to delve into the topics beyond narrow professional focus.

To my husband **Laurynas Sadauskas**, who did much more than babysitting when I was at conferences of finishing a paper before submission (a huge contribution in itself!). His wide intellectual curiosity and involvement inspired rich discussions, that led to the choice of my professional direction. His encouragement enabled me to follow this direction even through the darkest hours of my Ph.D. journey.

To my colleagues, who are the smartest and friendliest people that I know. Especially to **Karima Makhlouf** for her friendship and endless support. Also to my co-authors **Szilvia Lestyán**, **Carlos Pinzón**, **Héber H. Arcolezi**, **Kangsoo Jung** for the pleasure to work and have fun together.

To my mother **Rima Binkienė** for inspiring me as a child, when she was defending her own thesis in chemistry. To my father **Viktoras Binkis**, for inspiring me with his own example to not be afraid to change your path and seek your true passion. And to both of them for supporting me in all possible ways throughout my extended education years.

To my children **Bona Sadauskaitė** and **Teodoras Sadauskas** for accepting to wait "just a second", that lasts until the paper is submitted. I hope having witnessed my Ph.D. experience will someday ignite love for science in them. For now, they think "science" is a monster that chained their mom to her computer.

To the reviewers of the thesis, **Mario Fritz** and **Pascal Van Hentenryck**, for thorough and insightful reviews of my thesis. I deeply value your generous feedback and encouragement, and I am fortunate to have had the benefit of your expertise in reviewing my manuscript.

To the president of the jury **Stefan Haar** and the members of the jury **Miguel Couceiro**, **Sébastien Gambs** and **Daniele Gorla**, for dedicating their time and expertise to evaluate the thesis and for asking thought-provoking questions during the defense. I truly enjoyed the discussion that led to many new ideas and inspiration for future research.

A special thanks to **Miguel Couceiro** for a meticulous reading and keen eye for detail that helped to spot the errors and enhance the overall quality of the manuscript.

And finally, to all of my collaborators with whom I interacted during my doctoral studies.

This thesis is a testament to the collective effort, encouragement, and support of all these remarkable individuals. Thank you for being an integral part of this academic achievement.

Contents

List of Figures

List of Tables

Part I

Introduction

1

Ethical AI

Ethical [1] Artificial intelligence (AI) is a set of practices and theories within the field of AI and machine learning (ML) that abides by certain moral values shaping the potential impact of technology on human lives. The ethical principles that technologies are expected to satisfy may vary between the organizations and countries that formulated them. There are multiple versions of possible ethical guidelines for ethical AI. Some of the most important guidelines and regulations for AI include the European Research Council High-Level Expert Group [2] (2018, EU), European AI Act [3] (2022, EU), AI Bill of Rights [4] (2022, US), the Algorithmic Accountability Acct [5] (2022, US), deep synthesis provisions [6] (2023, China). Some of the common ethical AI principles include [1, 2]:

- fairness and non-discrimination,
- privacy,
- interpretability and explainability,
- safety and reliability.

In the next sections, we introduce the main subdomains of ethical AI research, based on these principles.

[1] Other terms include 'Trustworthy', 'Reliable', 'Human Centric'
[2] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[3] https://eur-lex.europa.eu
[4] https://www.whitehouse.gov/ostp/ai-bill-of-rights/
[5] https://www.congress.gov/bill/117th-congress/house-bill/6580/text
[6] https://www.china-briefing.com/news/china-to-regulate-deep-synthesis-deep-fake-technology-starting-january-2023/

1.1 Fairness

The social and legal stance against unfair discrimination against individuals and groups is one of the achievements of modern societies. Both EU and US laws foresee the protection of certain historically marginalized social identities such as race, ethnicity, gender, sexual orientation, and religious or political beliefs. In certain domains, such as hiring, the list extends to disability and age. The problem of discrimination is faced anew in the context of algorithmic decision-making. Despite stereotypic beliefs about algorithmic neutrality and objectivity, some bad-case examples have raised red flags. When scrutinized, predictions based on machine learning were found to be discriminatory against race or gender minorities.

One of the notorious examples of algorithmic discrimination is the 2016 ProPublica analysis of a COMPAS algorithm used in the American criminal justice system [4]. The algorithm was built to predict the risk of recidivism, which was taken into account when considering the release on bail or the length of a sentence. However, when making a mistake in prediction, the system was biased to falsely assign a lower risk score to white convicted criminals and a higher risk score to black ones.

Other examples include the case of discrimination against black patients in the healthcare need prediction system [5], lower predictive accuracy in face recognition for minority groups [6], gender discrimination in job advertising [7], sexist predictions in natural language processing [8] and others. Algorithmic discrimination is arguably more dangerous than previous historical instances of unfairness. The first reason is the scale of the decisions. Machine learning algorithms are capable of assigning millions of labels that affect millions of people in one day. The second danger of machine discrimination is the feedback loop of the machine learning cycle (Figure 1.1). The feedback loop is created when the data generated under the influence of unfair decisions is used for further training of the algorithm (or other algorithms). Such data becomes ever more biased, creating a self-perpetuating model of reality.

1.1.1 Sources of algorithmic discrimination

Discrimination in machine learning is dangerous; however, it is most often unintentional. It usually results from learning from the biased data in combination with the specifics of the learning algorithm. Discrimination can arise at any stage of the machine learning cycle (Figure 1.1). The studies by [9, 10] point out three main causes of algorithmic discrimination: bias in modeling, bias in training, and bias in usage. Bias in modeling is related to the choices of parameters or features. Bias in usage arises when predictions are misinterpreted or transferred to inappropriate contexts. Finally, bias in training is related to the data used for machine learning. Biased data is arguably the most prominent source of unfair algorithmic predictions. "Garbage in, garbage out" [7] has become a mantra in the data science community. Various data quality issues may arise due to data collection and data generation processes. Most types of bias relevant to algorithmic discrimination boil down to *representation* and *historical* biases.

[7] The phrase is usually attributed to Wilf Hey, a computer scientist at IBM [1].

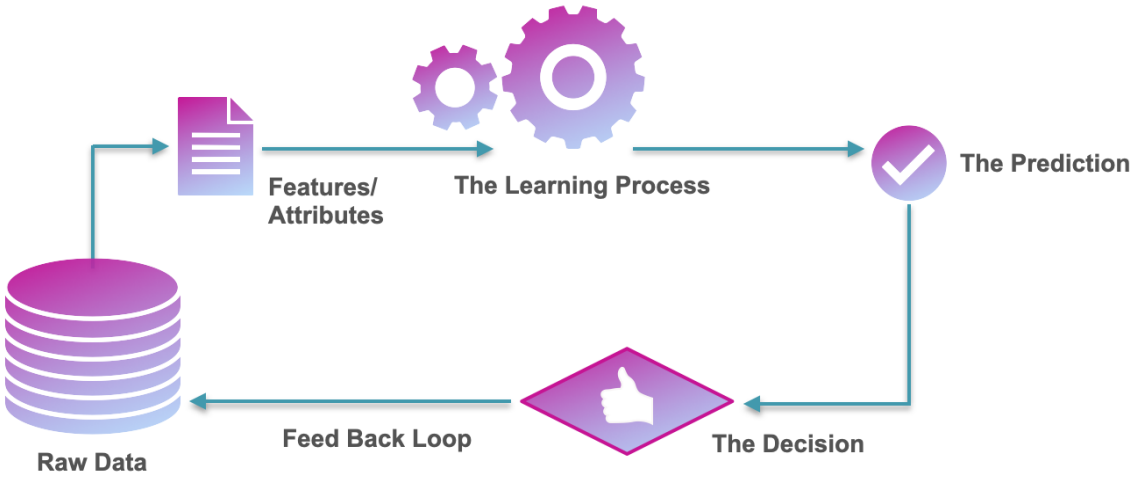


Figure 1.1: Machine learning cycle.

Historical bias

The traces of historical discrimination in the data may cause algorithmic unfairness. Discrimination can manifest itself as a lower positive decision rate for marginalized groups. For example, low hiring rates for eligible women could reflect discriminatory practices in the company. This is a direct discrimination recorded in the data set. There are also more subtle forms of historical bias that are produced by long-term structural inequalities in society. For example, data can have seemingly neutral labels but lower rates of merit-related qualities, such as university degrees, for the group that historically had less access to education. In this case, the algorithm might learn negative associations between a group and desirable qualities. One of the examples is the previously mentioned COMPAS algorithm used in the American criminal justice system [4]. Data used to train the COMPAS algorithm had higher baseline reoffense rates in the black population for historical reasons [11]. As a result, the algorithm overestimated the probability of reoffence for black defendants compared to white.

(Under)Representation of the group

Algorithmic discrimination can emerge from the training of an ML model using data with a disparity in the number of samples corresponding to each subpopulation of the sensitive group. The under-representation problem is well known in the deep learning literature. Buolamwini et al. [6] found that several commercial face recognition software had significantly lower accuracy for individuals belonging to a specific subpopulation, that is, dark-skinned women. The reason behind the disparity was the data set that was not representative of darker-skinned women. The authors of the paper were able to improve the fairness of the predictions by collecting more images of dark-skinned women. It should be noted that underrepresentation could also be the result of historical discrimination or marginalization. For example, there are fewer women profiles in the data on high-level managers. It could also be a sampling artifact. For example, collecting data on random passersby in summer will end up showing a higher number of tourists

in the city.

1.1.2 Fairness notions

The quantification of discrimination in data or algorithmic decisions uses formal expressions called *fairness notions*. There exist more than 40 different fairness notions. Fairness notions can be classified into three major categories: individual fairness [12, 13], group fairness [14, 15] and causal approaches [16, 17].

Individual fairness notions

Individual fairness notions are based on the idea that similar individuals should be treated similarly [13]. Although the idea is very intuitive, it is difficult to establish a measure of similarity and relevant attributes in which the individuals should be similar. For example, it does not make sense to take into account the similarity in the birthplace in the context of hiring. The similarity that matters could be related to work experience or education. Moreover, the segmentation of similar individuals can vary from fine-grained, which makes it computationally challenging, to very coarse, which approaches the group. Furthermore, individual fairness is not compatible with the disparate impact legal framework. The disparate impact framework imposes the equality of outcome between certain groups, even if the individuals between the groups are dissimilar [18]. The goal is to avoid a disproportionate negative impact on a group that was historically discriminated against or disadvantaged.

Group fairness notions

Group fairness notions consider groups that share the value of a sensitive attribute. They can be probabilistic or based on some measure of predictive performance. Probabilistic metrics measure the probabilities of positive labels conditioned on a sensitive attribute. Conditioning can also include ground truth labels or *explanatory* attributes that are supposed to justify a possible disparity. Fairness notions based on performance measurement usually require equality in confusion matrix measures (in classification tasks) or error rates (in regression) for sensitive groups. The main advantage of group fairness notions is that they are relatively easy to measure. However, they provide a meaningful assessment of fairness only in cases where the within-group distributions are fairly homogeneous.

Causal fairness notions

Causal fairness notions can be individual or group-based. Most causal metrics have statistical counterparts. The distinct feature of causal metrics is that they can distinguish discrimination from *causal biases* [8]. Causal biases can be introduced by the way data are collected or generated and can hinder accurate evaluation of discrimination by statistical measures. In extreme cases, such as Simpson's statistical paradox, the result measured by causal metrics can be the opposite of

[8] Causal biases are discussed in sections 10 and 11

that obtained by statistical evaluation [19]. A well-known example is a Berkley College admission case, in which discrimination against women was reversed to indicate discrimination against men when considering a causal approach to evaluating fairness [20, 21]. Causal fairness notions can also help to distinguish and quantify fair and unfair pathways in the data [22]. Namely, how much the sensitive attribute influences the decision through the *explanatory* variables (eg, education in hiring) that justify it, or the *red-lining* variables (eg. ZIP codes in hiring) that should not be used for the decision.

Application of fairness notions

The work of [23] summarizes the other criteria to favor a particular notion of fairness. One is the availability of ground truth in the case where the algorithmic decision is compared with the labels in the data. Here, the requirement is to have similar predictive power for sensitive groups. However, this approach assumes that the labels are reliable. Labels themselves could be biased by prejudiced human decisions or differential measurement errors for groups. If the labels can be trusted, then the specific measure for prediction power can be considered [23]. In some contexts, the false negative rate (how much of the predicted negatives are indeed positive) is more important than the false positive rate (how much or predicted positives are indeed negative). One example is cancer prediction, where a false negative would mean that a patient will not receive the treatment she needs, while a false positive would mean additional tests. In this case, it is meaningful to apply the fairness notion based on the false negatives rate. In the case of causal fairness notions, the applicability depends on the identifiability of causal quantities from the data [19]. However, it is important to note that the choice of a particular notion of fairness is highly subjective and depends on ethical values [24]. In this sense, it cannot be fully automated. Fairness notions encode cultural values and assumptions about the world. For example, should the distributions of positive decisions be equal for the sensitive groups, or disparity can be justified? Even if we find a justification for a disparity, it is highly context-dependent. For example, it is considered normal to favor one gender when hiring for a role in a movie or screening for a disease such as breast cancer. However, when hiring for an office job, gender should not play any role, unless one group on average has a more suitable education than the other. Here, "education" would be an explanatory factor that could justify the disparity in hiring decisions. However, what explanatory factor is suitable for a particular job, health screening, loan granting, or other decision is subject to the domain of application, cultural values, and legal framework.

1.1.3 Fairness correction methods

The mitigation of the bias in the data or algorithmic decisions can be done either before learning the model (pre-processing), at the time of learning (in-processing), or applied to already learned prediction labels (post-processing). There is no consensus as to which approach is superior to the others, and the applicability of a particular technique depends on the data set and the notion of fairness that it aims to satisfy [25]. The benefit of pre-processing techniques is that they

are model agnostic. However, pre-processing involves the manipulation of data that, if done without clearly stated assumptions, borders data falsification. In-processing has the advantage of explicitly incorporating the fairness-accuracy trade-off into the parameters and regularization terms of the algorithm. However, in most cases, it is specific to a particular algorithm. Finally, post-processing is a direct treatment of the labels assigned by the algorithm. It imposes the desired fairness properties but does so without optimizing the learning process itself.

1.2 Privacy

Protecting data privacy is a legal obligation in Europe and many other countries around the world. The first approach to protecting sensitive and identifying information in the data was data anonymization. This means that all personal identifiers, such as names or ID numbers, are removed from the data. However, there exist *quasiidentifiers* such as age or ZIP code that under certain conditions can lead to the reidentification of an individual in the database. Researchers have shown that data anonymization is not enough for privacy protection, because reidentification is possible by combining quasi-identifiers from multiple databases [26, 27]. To answer this need, numerous privatization methods have been developed to maximize the trade-off between a good level of data privacy and utility. Some examples of data privatization approaches include k-anonymity [28], t-closeness [29] and l-diversity [30]. In this thesis we will focus on one of the most popular data-privatization approaches - *Differential privacy* (DP) [31].

1.2.1 Differential Privacy

Differential privacy has gained popularity because of formal privacy guarantees and useful properties. DP is currently used in a variety of applications, from programming languages [32] to social networks [33] and geolocation [34]. The DP property establishes a bound on the ratio of the probability of getting the same reported answer from two adjacent databases, namely two databases that differ for just one record. In essence, DP provides guarantees that if any participant added or removed her data from the data set, no outputs would become significantly more or less likely [35]. The bound is expressed in terms of a parameter ϵ, which represents the level of privacy. The smaller the ϵ, the smaller the difference between databases with and without any particular record, thus the higher the level of privacy.

DP has composability and robustness to post-processing properties. Composability ensures that the joint distribution of the output of differentially private mechanisms also satisfies differential privacy. Robustness to post-processing guarantees that the output of any function applied to differentially private data will also be differentially private. Those properties are very useful for evaluating and ensuring privacy in the context of machine learning. DP can be further classified into *central* and *local*, depending on the point at which data privatization is applied (locally at the data owners' site or centrally at the data aggregator's site), and a metric variant called d-privacy.

Central DP

Central DP, which is the original notion of DP, assumes the existence of a trusted server where the data are aggregated. Data consumers (analysts) cannot access the data set directly but only query it via the server, which is supposed to obfuscate the answer by controlled noise, before reporting it to the analysts. One limitation of the central DP model is that the server or the data collector cannot always be trusted: they may collude with an attacker or just be unable to protect the data from security breaches.

Local DP

Local DP (LDP) has been proposed as an alternative to the central DP model [36, 37]. In LDP the individual data are obfuscated directly by the data provider before even being collected. The main advantage of LDP is that users are more willing to share their data when they do not need to rely on the trustworthiness of the data collector and the server. This model has become popular, especially because it has been adopted and promoted by leading high-tech companies such as Google [38], Microsoft [39], and Apple [40].

d-privacy

A variant of DP, called *d-privacy* (also known as *metric privacy*), was introduced in [41]. d-privacy is suitable for domains provided with a notion of distance. Like in central and local DP, d-privacy imposes a bound on the probability that the same result is obtained from two different objects (the arguments of the mechanism). However, unlike DP, this bound does not depend only on the parameter ϵ, but also on the distance between the objects. This means that the noise can be calibrated depending on how large the range in which we want to achieve indistinguishability is. On the contrary, LDP requires indistinguishability between any pair of elements in the domain. d-privacy, therefore, is particularly useful in applications where hiding an element within a group of neighbors is a sufficient measure of privacy protection. d-privacy has been applied especially in the local model, and in particular, in the context of location privacy, where it takes the name of *geo-indistinguishability* [42].

1.3 Causality

Causality is arguably a universal and intuitive way humans make sense of the world. It pervades many areas of human endeavor and is prominent in major religions. An attempt to systematically reflect on causal concepts goes back to Antiquity (Aristotle) [43]. The idea of causality was further developed by R.Descartes, I.Kant, G.W.Leibniz, and S.Hume [44]. One of the important milestones in the history of causality is the shift from a deterministic to a probabilistic understanding of the causal effect, which is strongly influenced by discoveries in quantum mechanics [45].

Causality has become fundamental for modern sciences, as well as ethics and legal philosophy and practice [44]. As a response, frameworks and tools have emerged to establish causal relationships from the data (or statistical causality). There exist two main approaches to causal

evidence: the analysis of (1) experimental data and (2) observational data. The first approach is based on randomized controlled trials (RCT), which is widely recognized as the gold standard for proving causal relationships [46–48]. However, random assignment is often impractical or impossible. For example, it is not ethical to assign random participants in the experiment to smoke or engage in other hazardous activities. It is also impossible to change someone's race or gender, to measure its impact on hiring decisions or income. This is why causality is often determined based on observable outcomes (the second approach). However, it requires the use of specific instruments to distinguish causality from statistical correlations.

The two most prominent frameworks for determining causality from observational data are the potential outcome framework [49], and the structural probabilistic model based on directed acyclic graphs (DAGs) [46]. Social and health sciences are dominated by the potential outcome framework [50], while DAGs are gaining popularity at the intersection of causality and AI [51].

1.3.1 Frameworks and Definitions

The realm of statistical causality is a mix of competing and sometimes complementary theories and concepts, rather than a single cohesive framework. The researchers in [52] compare the discipline of statistical causality with a "probability theory before Kolmogorov". In practice, statistical causality is applied using a combination of tools and approaches from several frameworks. We will provide a top-contour overview of the statistical causality landscape by introducing several existing theories and definitions of causality. Most of them rely on understanding causes as a relationship that is revealed by linear regression, grounds the definition of a cause in a notion of real or hypothetical intervention, or requires a mechanistic understanding of the underlying cause [53]. In this thesis, we rely mainly on the concepts and tools developed in the framework of structural probabilistic models [46] and the potential outcome [49] frameworks. Next, we provide more details on Pearl's and Rubin's theoretical background on causality. We also briefly mention other existing approaches and definitions of causality. The technical definitions of the relevant causal concepts can be found in the Technical Preliminaries 5.

Potential Outcome

The potential outcomes framework is one of the first formal theories of causal inference [54]. The framework defines causal effects as the difference of potential outcomes at different levels or the presence versus absence of exposure [49]. The language of potential outcomes allows one to express causal effects as statements about joint distributions of potential outcomes expressed as random variables. The causal assumptions are represented as restrictions on these distributions [55].

The potential outcomes can be *factual* - the outcome that happened, and *counterfactual* - the outcome that would have happened had the exposure been different. For example, if a person took a pill and got better, the factual outcome is "getting better". The counterfactual outcome is what would have happened had she not taken the pill. It is quite obvious that for this person we are not able to observe the counterfactual outcome. That makes the subject-specific effect of

limited practical use [54].

However, counterfactual outcomes and causal effects can be estimated at the population level. The population-level effect measures the aggregate impact of exposure on the outcome. The causal effect of exposure is conceptualized as a contrast between the outcome where everyone received treatment versus the outcome where no one received treatment. Once again, in the data, we observe only the factual outcomes at the factual levels of exposure received by each individual. However, given the SUTVA assumptions of identifiability [9] the causal effect can be estimated due to randomization [54]. If exposure is randomized, the potential outcomes are statistically independent of the exposure, and the conditional probability of the outcome, among those who factually received the treatment is equivalent to the one in which everyone received it. It is formally demonstrated in [49, 54].

The potential outcome framework provides tools for causal reasoning and simplified counterfactual computation [50]. However, it has shortcomings. Pearl criticizes the potential outcome framework for not providing clear rules for identifying relevant covariates [56]. He points out that including as many covariates as possible is a dangerous approach, because controlling for some covariates can increase the bias in the data.

Non-Parametric Structural Models (NPSEM)

Pearl [46] causality framework is often praised for its coherency and strong formal background [57]. Pearl synthesizes the approaches from agency causality (based on the concept of intervention) and probabilistic graphical models [57] as well as ideas from counterfactual or potential outcomes [54]. In terms of understanding the relationship between cause and effect, Pearl takes a middle ground between the purely stochastic causal Bayesian model approach (inspired by quantum mechanics) and the deterministic Laplace conception previously dominated in structural equation models (SEM) popular in econometrics and social sciences [46]. The NPSEM or Pearl framework links the graphical structure with the joint distributions of the variables. The causal structure is expressed by the directed acyclic graph (DAG). A DAG $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ is composed of a set of variables/nodes $\mathbf{V}$ and a set of directed edges $\mathcal{E}$ between them so that no cycle is formed. The DAG relates causal structure and joint distribution in the data through the Markov condition, where every variable is conditionally independent of its nondescendants given its parents. DAGs encode not only conditional independence relationships but also define the causal and non-causal data-generating processes. If node Y has an incoming arrow from node X, then X is the direct cause of Y. If Y has a direct arrow from variable Z, which has a direct arrow from X, then X is an indirect cause of Y (Z being a mediator). The confounder and collider structures imply a statistical relationship between the treatment and the outcome that is non-causal. Pearl also provides identifiability conditions that specify when and how causal quantities can be evaluated from observational data [46]. This is considered one of the most important contributions of the NPSEM framework [55].

[9] The assumptions are detailed in the Technical Preliminaries 5.

Other approaches

The sufficient cause framework conceptualizes causation as a set of sufficient causes that determine the occurrence of an event [58, 59]. Contrary to the potential outcomes or counterfactual approach, the main focus here is on the effect rather than the cause [58]. In particular, Pearl [46] has introduced alternative probabilistic notions of necessity and sufficiency [58]. The decision-theoretic approach assumes stochastic counterfactuals and focuses on the transportability of inference between an observational and an experimental regime [53]. It allows relaxation of strong assumptions required by the potential outcomes framework in some problems of causal inference [52]. Structural equation models (SEM) are deterministic models based on structural linear equations [60]. SEMs are limited in terms of their parametric assumptions and expressive power.

1.3.2 Causality and AI

The use of causality in AI falls mainly into one of two categories. The first approach is to employ artificial intelligence to enhance the qualitative discovery and/or quantification of causal connections from the data. The second one is to use causal tools to improve Machine Learning (ML) predictions. Next, we elaborate on both of these methods to combine causality and ML.

ML for causality

Causal Discovery Most of the techniques for obtaining causal quantities rely on knowing the causal structure of the data. It was previously assumed to be provided by experts. Recent advances in causal discovery offer algorithmic tools for recovering causal graphs from observational data. The basis for causal discovery is the probabilistic and graphical concepts of causality [57]. Two main groups of causal discovery algorithms can be distinguished based on their attempt to identify conditional or unconditional (including pairwise) independencies in the distribution from which the observational data is generated. The first category includes constraints and score-based algorithms such as PC [61], FCI [62], and GES [63]. They usually produce a partially oriented causal graph. The second category consists of algorithms based on causal asymmetries such as LiNGAM [64], and PNL [65]. The algorithms based on Kolmogorov's (algorithmic) complexity assume that if knowing the shortest compression of one variable does not reveal the shorter compression of the other, two variables are considered independent [51, 66]. The summary of the principles and performance for pairwise causal discovery is provided by Mooij et al. [67]. If the assumptions of the algorithms are satisfied, they are capable of identifying a unique causal graph or a causal direction between the two variables.

ML Tools for Causal Inference Supervised or semi-supervised machine learning methods can be used to estimate causal quantities from the data or for variable selection in situations with a high number of covariates [68, 69]. ML algorithms such as, for example, logistic regression, bagging, random forest, and others, can be beneficial in estimating propensity scores used to estimate causal effects in the potential outcome framework [70, 71].

Causality for ML

One of the main arguments that motivated the use of causality for machine learning is that causal modeling can lead to more invariant or robust models [51]. The problem of overfitting and vulnerability to a domain shift is a known problem in ML. It is intuitive that learning the correlation between two phenomena, for example, rain and umbrellas, will not help to predict rain in situations where people prefer raincoats instead of umbrellas. A causal understanding of phenomena is more general to multiple circumstances. Following Pearl, "...we may as well view our unsatiated quest for understanding how data is generated or how things work as a quest to acquire the ability to make predictions under a wider range of circumstances, including circumstances in which things are taken apart, reconfigured, or undergo spontaneous change" [46]. One of the methods to combine the ML model with the causal approach is to incorporate causal knowledge (usually in the form of a complete or partial causal graph) in the learning process [72, 73]. Causal representation learning is an attempt to combine latent variables derived from unstructured data and causal structure to arrive at a more invariant or fair model [51, 74–76]. The causal structure can also be used for feature selection, assuming that it is known. Models based on direct causes to predict the outcome are considered more robust [77].

1.4 Trade-Offs in Ethical AI and Causality

The ethical AI landscape, both as a whole and within specific disciplines, such as fairness or privacy, is marked by inherent trade-offs. Some of the most frequently encountered trade-offs include those between:

- fairness and accuracy [78–81];
- explainability and accuracy [82, 83];
- accuracy and generalizability [51];
- various fairness notions [24, 84, 85];
- privacy and accuracy [86, 87].

However, some ethical AI requirements can enhance another desired property. For example, the explainability of a model is important for evaluating fairness. There is also evidence that privacy is associated with robustness [88, 89]. Next, we will provide more details on the difficulty in achieving two or more ethical AI objectives simultaneously. We will also illustrate how causality can soften tensions and foster synergies among various aspects of ethical AI.

Most of the statistical fairness literature aims to improve some fairness metric while preserving accuracy as much as possible [90–93]. Often the level of achieved fairness is dependent on the willingness to make a sacrifice in accuracy. This loss in accuracy is a consequence of either obscuring information that is important for prediction but is also contributing to discrimination in the data or constraining the algorithm to produce prediction within certain boundaries of fairness. On the other hand, causality shifts attention from the accuracy with respect to the observed

labels to the outcomes based on causal knowledge or towards the more generalizable predictions. Causal fairness notions also allow one to re-evaluate discrimination by distinguishing justifiable and discriminatory paths between sensitive variables and the outcome or by identifying spurious correlations between them. As a consequence, the tensions between accuracy and fairness are less acute or, sometimes, eliminated. Causality also provides human-understandable explanations that facilitate reasoning about fairness. This is especially important when the judgment of fairness or discrimination is highly culturally subjective and should rely on a public consensus, rather than an algorithmic prediction or individual decision. This kind of judgment, of course, requires an understanding of the reasoning behind a prediction or decision. For example, what variables contribute to the lower rates of positive predictions? Many complex algorithms, such as deep neural networks (DNN) or random forest (RF), have impressive predictive power but provide "black-box" solutions that are hard to question or evaluate. Causal models are inherently explainable because they often rely on an explicit causal structure or well-articulated causal assumptions. As an added value, causal models provide better generalizability and precision in the presence of a shift in distributions [51]. Statistical learning is vulnerable to relying on spurious correlations in the data. One known example is a computer vision algorithm learning to recognize cows based on the association with grass in the background [94]. As a consequence, it no longer recognizes a cow if it is located on a beach. In contrast, the model based on causal features can produce accurate predictions despite changes in the data distribution.

Fairness is also subject to tensions between different notions and discrimination metrics. Numerous approaches to address the quantification of fairness suffer from conceptual and mathematical disagreement. Friedler et al. [24] point out the worldview incompatibility between the "what you see is what you get" and "we are all equal" conceptual frameworks. The first framework justifies disparity since it is in line with the ground-truth labels or explanatory variables in the data corresponding to a certain measurement for merit or need. For example, the hiring rate for two groups may be different since the level of education is also different. On the contrary, the "we are all equal" framework would require equality in hiring rates because any differences in education are due to the historical structural inequalities that should be compensated for and erased from future decisions. In this case, causality helps to better articulate the adopted fairness approach. It requires evaluating the data-generating process and the pathways between the sensitive attribute and the outcome. This makes the assumptions behind the choice of a metric explicit.

Kim et al. [85] provide a formalism for systematic reasoning about group fairness notions by expressing them as functions of the fairness–confusion tensor. The authors prove the general incompatibility between multiple notions of fairness and provide the necessary conditions under which they can be satisfied. However, this incompatibility does not hold for fairness metrics under the causal framework [95].

The differential privacy approach relies on adding noise to the data which is controlled by the parameter ϵ (the smaller value of ϵ corresponds to more noise, while the larger value indicates less noise and less privacy). Naturally, it hurts the accuracy of an algorithm learned on the privatized data. It is yet unknown how to avoid this fundamental trade-off between data

protection and the utility of the data. However, some results show that causal models can be more robust to membership inference attacks at a lower value of ϵ (thus less noise) [77].

2

Goals and Contributions

In this thesis, we explicitly focus on two principles of ethical AI, fairness and privacy. We investigate the relationship between fairness, privacy, and causality. In addition, we explore the role of causality to implicitly enhance interpretability and explainability of AI models.

2.1 Contributions

The general contribution of this thesis includes the holistic approach to Ethical AI, which aims to incorporate methods for fairness, privacy, and causality. We analyze and propose methods to accurately measure bias in the data. We distinguish and formalize sources of discrimination such as sample size and under-representation biases and causal biases that can hinder accurate measurement of direct discrimination. We discuss the need and challenges of applying causality in ML fairness from a statistical and legal point of view. We identify a causal graph as one of the main requirements for using causality in ethical AI. Motivated by this, we explore the impact of data privatization on the learnability of causal graphs from the data. We provide insights on achieving better trade-off in causal discovery and local data privacy. Finally, we propose a causal knowledge-based data debiasing approach that reconciles fairness, accuracy, and explainability in algorithmic decision-making. Next, we detail the contributions by chapter (the original contributions of the thesis start at chapter 7):

Chapter 7: BaBE: Enhancing Fairness via Estimation of Latent Explaining Variables

Our first significant contribution lies in the proposal of a novel approach to estimate the conditional distribution of a latent explaining variable $E|S$, a method that utilizes the Expectation-Maximization (EM) technique. This estimation serves as a foundation for pre-processing data to

improve fairness and increase accuracy (with respect to the decision based on true explaining variable) and explainability of the decision.

Chapter 8: Underrepresentation and Sampling Bias in Machine Learning

We systematically analyze the impact of sample size and underrepresentation of a group on discrimination in algorithmic decisions. We link our analysis to bias mitigation techniques and show that, in the presence of underrepresentation bias, collecting more data samples for the underrepresented group typically amplifies discrimination rather than reducing it.

Chapter 9: Causal Discovery under Local Privacy

In the domain of privacy, our work systematically compares the performance of different locally private mechanisms, specifically Geometric and k-RR, in the context of causal discovery tasks. With our findings, we highlight the advantages of using geometric obfuscation methods over k-RR, shedding light on the impact of noise levels on algorithm performance.

Chapter 10: On the Need and Applicability of Causality for Fair Machine Learning

We emphasize the necessity of incorporating causality into fair AI, consolidating statistical and legal arguments. Compared to existing work, ours is the first attempt to connect causality in fair AI with the European AI legislation. In addition, we discuss the requirements of applying causality to fairness evaluation in practice.

Chapter 11: Dissecting Causal Biases

We explore causal biases in relation to the precise measurement of the direct effect of group membership on the outcome. Leveraging tools from the field of causality, we develop closed-form expressions for various sources of causal biases, including confounding, colliding, and measurement. In addition, we introduce interaction bias, which is not previously discussed in the context of fairness. Our empirical analysis highlights the extent of causal biases in fairness benchmark datasets, underlining the importance of addressing these biases in machine learning.

Chapter 12: Gender and Sex Bias in COVID-19 Data

In the context of COVID-19, we contribute a comprehensive review of the literature, highlighting the greater vulnerability of men than women to the virus. We use causal graphs to analyze the causal relationship between gender and COVID-19, perform causal analysis on synthetic data, and underscore the importance of explainability and causality in understanding big data and facilitating equitable data-driven decisions.

Together, our contributions advance the field of ethical AI by drawing pathways toward better trade-offs between fairness, accuracy, privacy, and explainability.

2.2 List of Publications

The content of this dissertation is based on the following publications. The bibliographical information and abstracts are listed below.

2.2.1 Publications included in this thesis

Part III (Fairness) is based on:

1. Binkyte, R., Gorla, D. and Palamidessi, C., 2023. BaBE: Enhancing Fairness via Estimation of Latent Explaining Variables. arXiv preprint arXiv:2307.02891. *Submitted*.

We consider the problem of unfair discrimination between two groups and propose a pre-processing method to achieve fairness. Corrective methods such as statistical parity usually lead to bad accuracy and do not really achieve fairness in situations where there is a correlation between the sensitive attribute S and the legitimate attribute E (explanatory variable) that should determine the decision. To overcome these drawbacks, other notions of fairness have been proposed, in particular, conditional statistical parity and equal opportunity. However, E is often not directly observable in the data, i.e., it is a latent variable. We may observe some other variable Z representing E, but the problem is that Z may also be affected by S, hence Z itself can be biased. To deal with this problem, we propose BaBE (Bayesian Bias Elimination), an approach based on a combination of Bayesian inference and the Expectation-Maximization method, to estimate the most likely value of E for a given Z for each group. The decision can then be based directly on the estimated E. We show, by experiments on synthetic and real data sets, that our approach provides a good level of fairness as well as high accuracy.

2. Zhioua, S. and Binkytė, R., 2023. Shedding Light on Underrepresentation and Sampling Bias in Machine Learning. arXiv preprint arXiv:2306.05068. *Submitted*

Accurately measuring discrimination is crucial to faithfully assess the fairness of trained machine learning (ML) models. Furthermore, understanding the bias responsible for discrimination in the data guides the appropriate mitigation approach. Sampling bias is one of the common ML biases, which, in case where it is born differently by different groups (e.g. females vs males, whites vs blacks, etc.), may exacerbate discrimination against specific subpopulations. However, despite its familiarity, sampling bias is not well defined and is inconsistently used in the literature. In this paper, we attempt to disambiguate this term by introducing clearly defined variants of sampling bias, namely, sample size bias (SSB) and underrepresentation bias (URB). We also show how discrimination can be decomposed into variance, bias, and noise. Finally, we challenge the commonly accepted mitigation approach that discrimination can be addressed by collecting more samples of the underrepresented group.

Part IV (Privacy) is based on:

3. Binkytė, R., Pinzón, C., Lestyán, S., Jung, K, Arcolezi, H., Palamidessi, C. 2023.Causal Discovery under Local Privacy. *Accepted at Causal Learning and Reasoning (CLeaR) 2024*

***Conference*.**

Differential privacy is a widely adopted framework designed to protect sensitive data providers within a data set. It is based on the application of controlled noise at the interface between the server that stores and processes the data, and the data consumers. Local differential privacy is a variant that allows data providers to apply the privatization mechanism themselves to their data individually. Therefore it provides protection also in contexts in which the server, or even the data collector, cannot be trusted. However, the introduction of noise inevitably affects the utility of the data, particularly by distorting the correlations between individual data components. This distortion can be detrimental to tasks such as causal discovery. In this paper, we consider various well-known locally differentially private mechanisms and compare the trade-off between the privacy they provide, and the accuracy of the causal structure produced by algorithms for causal learning when applied to data obfuscated by these mechanisms. Our analysis yields valuable insights into selecting appropriate local differentially private protocols for causal discovery tasks. We foresee that our findings will aid researchers and practitioners in conducting locally private causal discovery.

Part V (Causality) is based on:

4. Binkytė, R., Grozdanovski, L. and Zhioua, S. 2023. On the Need and Applicability of Causality for Fair Machine Learning. arXiv preprint arXiv:2207.04053. *Submitted*.

Besides its common use cases in epidemiology, political and social sciences, causality turns out to be crucial in evaluating the fairness of automated decisions, both in a legal and everyday sense. We provide arguments and examples of why causality is particularly important for fairness evaluation. In particular, we point out the social impact of non-causal predictions and the legal anti-discrimination process that relies on causal claims. We conclude with a discussion of the challenges and limitations of applying causality in practical scenarios, as well as possible solutions.

5. Binkytė, R., Zhioua, S. and Turki, Y., 2023. Dissecting Causal Biases. arXiv preprint arXiv:2310.13364. *Submitted*.

Accurately measuring discrimination in machine learning-based automated decision systems is required to address the vital issue of fairness between subpopulations and/or individuals. Any bias in measuring discrimination can lead to either amplification or underestimation of the true value of discrimination. This paper focuses on a class of bias originating in the way training data is generated and/or collected. We call such class causal biases and use tools from the field of causality to formally define and analyze such biases. Four sources of bias are considered, namely, confounding, selection, measurement, and interaction. The main contribution of this paper is to provide, for each source of bias, a closed-form expression in terms of the model parameters. This makes it possible to analyze the behavior of each source of bias, in particular, in which cases they are absent and in which other cases they are maximized. We hope that the provided characterizations help the community better understand the sources of bias in machine learning applications.

6. Díaz-Rodríguez, N., Binkytė, R., Bakkali, W., Bookseller, S., Tubaro, P., Bacevičius, A., Zhioua, S. and Chatila, R., 2023. Gender and Sex Bias in COVID-19 Epidemiological Data through the Lens of Causality. Information Processing and Management, 60(3), p.103276.

The COVID-19 pandemic has spurred a large number of experimental and observational studies that report a clear correlation between the risk of developing severe COVID-19 (or dying from it) and whether the individual is male or female. This paper is an attempt to explain the supposed male vulnerability to COVID-19 using a causal approach. We proceed by identifying a set of confounding and mediating factors, based on the review of the epidemiological literature and the analysis of sex-disaggregated data. Those factors are then taken into account to produce fair and explainable prediction and decision models from observational data. The paper outlines how non-causal models can motivate discriminatory policies such as biased allocation of the limited resources in intensive care units (ICUs). The objective is to anticipate and avoid disparate impact and discrimination, by considering causal knowledge and causal-based techniques to complement the collection and analysis of observational big data. The hope is to contribute to the more careful use of health-related information access systems to develop fair and robust predictive models.

2.2.2 Other publications

Other works that were published during my Ph.D. which are not included in the thesis content:

7. Binkytė, R., Makhlouf, K., Pinzón, C., Zhioua, S. and Palamidessi, C., 2023, June. Causal discovery for fairness. In NeurIPS Workshop on Algorithmic Fairness through the Lens of Causality and Privacy (pp. 7-22). PMLR.

It is crucial to consider the social and ethical consequences of AI- and ML-based decisions for the safe and acceptable use of these emerging technologies. Fairness, in particular, guarantees that ML decisions do not result in discrimination against individuals or minorities. Identifying and measuring reliably fairness/discrimination is better achieved using causality which considers the causal relation, beyond mere association, between the sensitive attribute (e.g. gender, race, religion, etc.) and the decision (e.g. job hiring, loan granting, etc.). The big impediment to the use of causality to address fairness, however, is the unavailability of the causal model (typically represented as a causal graph). Existing causal approaches to fairness in the literature do not address this problem and assume that the causal model is available. In this paper, we do not make such an assumption, and we review the major algorithms for discovering causal relations from observable data. This study focuses on causal discovery and its impact on fairness. In particular, we show how different causal discovery approaches may result in different causal models, and, most importantly, how even slight differences between causal models can have a significant impact on fairness/discrimination conclusions. These results are consolidated by empirical analysis using synthetic and standard fairness benchmark datasets. The main objective of this study is to highlight the importance of the causal discovery step to appropriately address fairness using causality.

8. Binkyte, R., 2023. Distant Reading and Viewing:“Big Questions” in Digital Art History and Digital Literary Studies. Digital Humanities Quarterly, 17(2).

The emergence of digital art history is influenced by advances in computer vision, on the one hand, and digitization of visual image archives - Getty Research Institute, Google Art Project, on the other. However, the quantitative approach to images has qualitative implications. Here, we explore how art history can be enriched with approaches that consciously apply computational methods together with theory and interpretation. We draw inspiration and examples from the earlier adoption of large-scale digital text analysis in digital literary studies.

Part II

Technical Preliminaries

3 Fairness

3.1 Fairness notions

Fairness metrics are quantitative measures used to assess and quantify the fairness of algorithms, models, or decision-making processes. They provide a way to evaluate and compare the performance of these systems in terms of how they treat different groups or individuals. We denote the sensitive attribute S, the decision or label in the data Y Pand the algorithmic prediction $\hat{Y}$. Many fairness metrics are defined in the context of binary classification, where the goal is to predict one of two classes, typically denoted as positive $Y = 1$ or negative $Y = 0$.

Confusion Matrix

To define fairness metrics, the confusion matrix is often used. Confusion matrix is a 2x2 table that summarizes the model's performance. It includes the following components:

- True Positives (TP): The number of positive instances correctly classified as positive.
- False Positives (FP): The number of negative instances incorrectly classified as positive.
- True Negatives (TN): The number of negative instances correctly classified as negative.
- False Negatives (FN): The number of positive instances incorrectly classified as negative.

The confusion matrix allows us to calculate various fairness metrics based on these components.

Statistical parity difference (SPD)

Statistical parity difference measures [96] the difference in the probability of favorable outcomes for the protected group and the nonprotected group. It is defined as:

DEFINITION 3.1.1.

$$\mathbb{P}[Y = 1|S = 1] - \mathbb{P}[Y = 1|S = 0] \tag{3.1}$$

Where Pr(Y=1 | S=1) is the probability of a positive outcome for the protected group. Pr(Y=1 | S=0) is the probability of a positive outcome for the unprotected group. A value equal to zero indicates no discrimination.

Disparate Impact (DI)

Disparate Impact [90] is similar to statistical parity difference, except that it measures the ratio of favorable outcomes for the protected group to the nonprotected group. It is defined as:

DEFINITION 3.1.2.

$$DI = \frac{\mathrm{P(Y=1|S=1)}}{\mathrm{P(Y=1|S=0)}}$$

Where Pr(Y=1|S=1) is the probability of a positive outcome for the protected group. Pr(Y=1|S=0) is the probability of a positive outcome for the non-protected group.

A value of DI significantly different from 1 indicates potential discrimination.

Equal Opportunity Difference (EOD)

Equal Opportunity Difference [97] assesses whether the correct positive predictions are the same for different groups. It is defined as:

DEFINITION 3.1.3.

$$EOD = \mathbb{P}[\hat{Y} = 1|Y = 1, S = 1] - \mathbb{P}[\hat{Y} = 1|Y = 1, S = 0] \tag{3.2}$$

Where $P[\hat{Y} = 1|Y = 1, S = 1]$ is the probability of correct positive prediction for privileged group and $P[\hat{Y} = 1|Y = 1, S = 0]$ - for the protected group. Equal Opportunity Difference should ideally be 0 for fairness.

Conditional Statistical Parity Difference (CSPD)

The conditional statistical parity difference [98] measures the fairness of the prediction based on an explanatory attribute E, which justifies the disparity. It is defined as:

DEFINITION 3.1.4.

$$\begin{aligned} CSPD &= \mathbb{P}[\hat{Y} = 1|E, S = 1] - \mathbb{P}[\hat{Y} = 1|E, S = 0] \\ &= \sum_{E} \mathbb{P}[\hat{Y} = 1|E = e, S = 1]\mathbb{P}[E = e] \\ &- \sum_{E} \mathbb{P}[\hat{Y} = 1|E = e, S = 0]\mathbb{P}[E = e] \end{aligned} \tag{3.3}$$

Where $\mathbb{P}[\hat{Y} = 1|E = e, S = 1]$ is the probability of positive prediction based on the value of E for $S = 1$ and $\mathbb{P}[\hat{Y} = 1|E = e, S = 0]$ - for $S = 0$. The conditional statistical parity difference equal to zero indicates fairness.

These equations provide a foundation for understanding and quantifying fairness metrics in machine learning, especially in binary classification scenarios. Keep in mind that fairness metrics can vary based on the specific goals and requirements of a given application, and different fairness definitions may involve different equations and calculations.

3.2 Bias mitigation

Algorithmic bias mitigation, also known as bias reduction or debiasing, refers to the process of identifying, reducing, or eliminating bias in the outcomes and decisions made by machine learning algorithms and models. Depending on the phase of the learning cycle at which the debiasing is applied, the techniques can be classified into pre-processing, in-processing, and post-processing. Next, we provide examples of each data debiasing approach.

3.2.1 Pre-processing

Fairness pre-processing techniques are strategies and methods applied to the training data before building machine learning models to reduce bias and ensure fairness. These techniques aim to address disparities in data with respect to sensitive attributes, such as gender, race, or age.

Re-sampling

Re-sampling [99] involves modifying the training dataset to balance the representation of different groups or classes, particularly those that are underrepresented. Techniques for over-sampling the minority class or under-sampling the majority class are applied to achieve a more balanced distribution.

Re-weighting

Re-weighting [100] data instances adjusts the importance or influence of each instance based on its group membership, ensuring that underrepresented groups have more weight. The data are pre-processed by assigning higher weights to instances from underrepresented groups.

Disparate Impact Remover

The Disparate Impact Remover [90] is a pre-processing technique that modifies the training data by equalizing the distributions of features for sensitive attributes to ensure that they do not disproportionately affect the model's decisions.

Data Transformation

Data transformation [101] techniques modify the data to reduce bias by adjusting the distribution of sensitive attributes. The data are transformed so that sensitive attributes exhibit similar distributions across groups.

Massaging

Massaging [102] involves modifying the data set by adding or removing instances to achieve fairness objectives. Data is debiased by adding instances to underrepresented groups or removing instances that may introduce bias.

3.2.2 In-processing

Fairness in-processing techniques involve integrating fairness considerations directly into the machine learning model during the training process. These techniques aim to reduce bias and promote fairness by adjusting the behavior of the model and predictions based on sensitive attributes.

Adversarial debiasing

Adversarial networks [103] are added to the model to learn a representation of the data that removes the influence of sensitive attributes. The model is trained to simultaneously optimize prediction accuracy and fairness, making it difficult to determine sensitive attributes from the data.

Regularization

In regularization [14] technique, fairness constraints are introduced during model training to minimize the impact of sensitive attributes. Regularization terms are added to the loss function to penalize models for making biased predictions.

3.2.3 Post-processing

Fairness post-processing techniques are strategies applied after a machine learning model has been trained to mitigate bias and ensure fairness in the model's predictions. These techniques adjust the model's outputs or decisions to align them with fairness objectives.

Re-ranking

Re-ranking [104] involves changing the order of model predictions based on fairness criteria to ensure that sensitive groups are not unfairly disadvantaged.

Calibration Post-processing

Calibration post-processing [105] techniques adjust model predictions to achieve calibrated probabilities, where the predicted probabilities reflect the true likelihood of an event occurring.

4 Privacy

Data privacy, often referred to as information privacy or data protection, is a fundamental concept that addresses the management, handling, and protection of personal and sensitive information. It concerns the rights and expectations of individuals regarding how their data is collected, processed, stored, and shared. Data privacy is a critical aspect of data security and ethical data handling, particularly in the digital age where vast amounts of personal information are collected and processed.

4.1 Privacy Notions

4.1.1 Central Differential Privacy

Central differential privacy (DP) [106] is a privacy framework used in the context of data analysis and data release. It ensures that individual-level data is protected by adding noise or randomization to the data prior to any analysis. Central Differential Privacy guarantees that the privacy of individuals is preserved while still allowing meaningful aggregate analyses to be performed on the data. Central DP assumes a trusted data aggregator.

Central DP ensures that the output of a data analysis algorithm, such as statistical queries or machine learning models, remains statistically indistinguishable when any single individual's data is included or excluded from the database.

Definition 4.1.1 (Differential Privacy). A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|X|}$ is (ϵ, δ)-differentially private if for all $\mathcal{S} \subseteq Range(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|X|}$

such that $|x - y|_1 \leq 1$:

$$\mathbb{P}[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon) Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta, \tag{4.1}$$

If $\delta = 0$, we say that $\mathcal{M}$ is ϵ-differentially private.

4.1.2 Local Differential Privacy

Local Differential Privacy is a privacy-preserving framework that focuses on protecting the privacy of individual data points in a data set while still enabling aggregate analyses and data-driven decision-making. Unlike Central Differential Privacy, which adds noise to the final output, Local Differential Privacy injects noise or randomization directly into individual data points before any analysis is performed. This ensures that the privacy of each individual's data is safeguarded, making it well-suited for scenarios where the data is highly sensitive and individual-level privacy is a top priority.

One of the widely used privacy models is LDP [36, 37], which is formally defined as follows.

> **Definition 4.1.2** (ϵ-Local Differential Privacy)**.** Let $\epsilon > 0$ be a parameter representing the level of privacy loss. A randomized mechanism $\mathcal{M}$ satisfies ϵ-local-differential-privacy (ϵ-LDP) if, for any pair of input values $v_1, v_2 \in Domain(\mathcal{M})$, and any possible output x of $\mathcal{M}$, the following holds (where $\mathbb{P}[e]$ represents the probability of the event e):
>
> $$\mathbb{P}[\mathcal{M}(v_1) = x] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(v_2) = x] \ .$$

In essence, LDP guarantees that it is unlikely that the data aggregator infers the true value from the reported data. Privacy loss ϵ controls the trade-off between privacy and utility. Note that lower values of ϵ result in tighter privacy protection. Similarly to global DP, LDP also has several fundamental properties, such as robustness to post-processing and composition [106].

4.1.3 Local d-Privacy

d-Privacy [41] is a variant of differential privacy that is particularly suitable for the domains that have a notion of distance. In essence, in the local model d-Privacy guarantees, like in LDP, that it is unlikely that the data aggregator or an attacker will infer the true value v from the reported data. But in this case, it is because it is made indistinguishable from all the other values in the neighborhood. In other words, the nearby secrets should look identical to any observer.

d-Privacy assumes that the domain of the mechanism $\mathcal{M}$ is provided with a notion of distance d.

> **Definition 4.1.3** (Local d-Privacy)**.** A mechanism $\mathcal{M}$ satisfies d-privacy, with privacy parameter ϵ, iff for all values, $v_1, v_2 \in Domain(\mathcal{M})$ and all possible outputs

x, the following inequality holds:

$$\mathbb{P}[\mathcal{M}(v_1) = x] \leq e^{\epsilon\, d(v_1, v_2)} \cdot \mathbb{P}[\mathcal{M}(v_2) = x] \ .$$

One of the best-known applications of d-privacy is in the context of location privacy, where it takes the name of *geo-indistinguishability* [42].

5 Causality

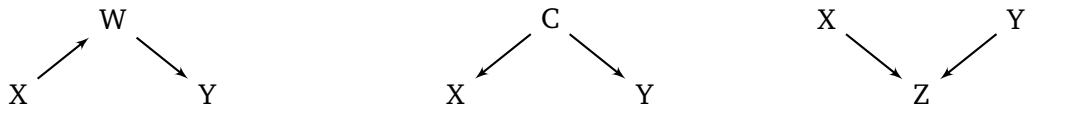
5.1 Causal structures

Variables are denoted by capital letters (e.g. X, Y). Small letters denote specific values of variables (e.g., $A = a$, $W = w$). Bold capital (e.g. $\mathbf{V}$) and small letters (e.g. $\mathbf{v}$) denote a set of variables and a set of values, respectively.

A causal graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, composed of a set of variables/vertices $\mathbf{V}$ and a set of edges $\mathcal{E}$, is a directed acyclic graph (DAG) that describes the causal relations between variables. Edges have causal interpretations. That is, a directed edge $X \rightarrow Y$ indicates a causal relation from the cause variable X to the effect variable Y. Consequently, if all other variables are fixed to some values and we change the value of X, Y will change, but not the other way around (changing the value of Y will not change the value of X).

There are three basic structures in a causal graph, namely, a mediator, a confounder, and a collider [21]. Figure 5.1 shows an example of each of these structures. The variable W in Figure 5.1(a) is called a mediator because it mediates the causal effect of X on Y[1]. A confounder variable (C is a common cause of two other variables (X and Y)). It is important to mention that in both mediator and confounding structures, X and Y are correlated. The difference is that in a mediator, X is a cause of Y, but in a confounder, X is not a cause of Y. They are simply correlated. A collider, on the other hand, is a variable caused by two other variables (Z in Figure 5.1(c))[2]. Unlike the two other structures, in the presence of a collider, X and Y are not correlated. However, if we condition on Z, X and Y become correlated.

[1] The mediator structure is known also as chain structure.
[2] A collider structure is known also as v-structure.



((a)) Mediator ((b)) Confounder ((c)) Collider

Figure 5.1: Basic structures of causal graphs.

As the causal relation between two variables can go through different paths, mediation analysis consists in distinguishing these causal paths. For example, in Figure 5.2, a causal effect between X and Y can be divided into direct ($X \to Y$), indirect ($X \to R \to Y$ and $X \to E \to Y$), or path-specific effect (e.g. only $X \to E \to Y$). Assuming X is a sensitive variable (used for discrimination), this is very relevant to fairness as a direct effect is always unfair because the sensitive variable should not be used directly to decide about the outcome, while the indirect or path-specific effects may be unfair or fair depending on the mediator variable: an indirect effect through a redlining/proxy variable (R) is unfair, while an indirect effect through an explaining variable (E) is acceptable (fair). A proxy variable is a descendent of X that is significantly correlated with it in such a way that the use of the proxy in the outcome Y has almost the same impact as using X directly. An explaining variable is also a descendent of X used to decide about the outcome Y that is influenced by X in a way that is accepted as non-discriminatory. For example, discrimination against women for job hiring is acceptable if it is justified by the low education level of female candidates. Deciding if a mediator is a proxy or explaining variable typically requires some expertise about the context of the problem.

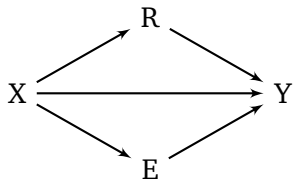


Figure 5.2: The causal effect between X and Y can be split into three different paths: direct ($X \to Y$) and indirect ($X \to R \to Y$ and $X \to E \to Y$), involving (R)edlining/proxy and (E)xplanatory variables.

Using causality allows one to appropriately assess fairness (and consequently discrimination) due to two main reasons. First, by identifying confounder variables between X and Y, it becomes possible to account for the noncausal effect that goes through the confounder variables. For example, the effect going through the path $X \leftarrow C \to Y$ in Figure 5.1(b) is non-causal while all paths between X and Y in Figure 5.2 correspond to causal effects. This is the reason we say that "causation is different than correlation." Second, causal mediation analysis allows us to split the total causal effect of X to Y into direct/indirect and fair/discriminatory effects.

5.2 Intervention and *do*-operator

The causal effect between two variables is typically expressed in terms of intervention probabilities. Intervention, noted $do(V = v)$ [46], is a manipulation of the model that consists in fixing the

value of a variable (or a set of variables) to a specific value regardless of the causes of that variable. The intervention $do(V = v)$ induces a different distribution on the other variables. Intuitively, while $\mathbb{P}(Y|A = a)$ reflects the population distribution of Y among individuals whose A value is a, $\mathbb{P}(Y|do(A = a)$[3] reflects the population distribution of Y if *everyone in the population* had their A value fixed at a. The obtained distribution $\mathbb{P}(Y|do(A = a)$ can be considered as a *counterfactual* distribution since the intervention forces a to take a value different from the one it would take in the actual world. $\mathbb{P}(Y|do(A = a)$ is not always computable from the data, a problem known as identifiability. For instance, if all counfounder variables are observable, the intervention probability, $\mathbb{P}(Y|do(A = a))$, can be computed by adjusting on the counfounder(s). For instance, assuming Z is the only confounder of A and Y, a back door formula (Equation 11.1) can be applied.

5.3 Causal fairness notions

Total Effect

Total effect (TE) [21][4] is the causal version of SPD (Equation 7.1) and is defined in terms of experimental probabilities as follows:

Definition 5.3.1.

$$TE_{x_1,x_0}(y) = \mathbb{P}(Y = y|do(X = x_1)) - \mathbb{P}(Y = y|do(X = x_0)) \tag{5.1}$$

TE measures the effect of the change of X from x_0 to x_1 on $Y = y$ along all the causal paths from X to Y. Intuitively, while SPD 3.1 reflects the difference in proportions of $Y = y$ in the current cohort, TE reflects the difference in proportions of $Y = y$ in the entire population. $\mathbb{P}(Y = y|do(X = x))$ denotes the probability of $Y = y$ after an intervention $do(X = x)$. This is equivalent to the probability of $Y = y$ after forcing all individuals in the population to have a value $X = x$. $\mathbb{P}(Y = y|do(X = x))$ is denoted $\mathbb{P}(y_x)$ for short[5].

Mediation analysis related notions

Mediation analysis is about distinguishing the different paths of the causal effect between two variables X and Y. Causal paths can be direct or indirect. The direct natural effect (NDE) [107] is the simplest notion of mediation analysis, which measures the direct causal effect between two variables. (e.g. X and Y). Assuming the variable X is binary (it can take two possible values x_0 and x_1), NDE is defined as:

Definition 5.3.2.

$$NDE_{x_1,x_0}(y) = \mathbb{P}(y_{x_1,\mathbf{z}_{x_0}}) - \mathbb{P}(y_{x_0}) \tag{5.2}$$

[3]The notations $Y_{A\leftarrow a}$ and $Y(a)$ are used in the literature as well. $\mathbb{P}(Y = y|do(A = a)) = \mathbb{P}(Y_{A=a} = y) = \mathbb{P}(Y_a = y) = \mathbb{P}(y_a)$ is used to define the causal effect of A on Y.

[4]Total Effect is also known as average causal effect (ACE).

[5]The notations $Y_{X\leftarrow x}$ and $Y(x)$ are used in the literature as well. $\mathbb{P}(Y = y|do(X = x)) = \mathbb{P}(Y_{X=x} = y) = \mathbb{P}(Y_x = y) = \mathbb{P}(y_x)$ is used to define the causal effect of X on Y.

Where $\mathbf{Z}$ is the set of mediator variables and $\mathbb{P}(y_{x_1,\mathbf{z}_{x_0}})$ is the probability of $Y = y$ had X been x_1 and had $\mathbf{Z}$ been the value it would naturally take if $X = x_0$. Using the graph in Figure 5.2, this means that X is set to x_1 in the single direct path $X \rightarrow Y$ (there is always only one direct path but several indirect paths between X and Y) and is set to x_0 in all other indirect paths ($X \rightarrow R \rightarrow Y$ and $X \rightarrow E \rightarrow Y$).

The natural indirect effect (NIE) [107] measures the indirect effect of X on Y and is defined as:

DEFINITION 5.3.3.

$$NIE_{x_1,x_0}(y) = \mathbb{P}(y_{x_0,\mathbf{z}_{x_1}}) - \mathbb{P}(y_{x_0}) \tag{5.3}$$

Using the same graph (Figure 5.2), this means that X is set to x_0 in the single direct path $X \rightarrow Y$ and is set to x_1 in all other indirect paths. The problem with NIE is that it does not distinguish between the fair (explainable) and unfair (indirect discrimination) effects.

The path-specific effect [21, 108, 109] is a more nuanced measure that characterizes the causal effect in terms of specific paths. Given a path set π, the π-specific effect is defined as:

DEFINITION 5.3.4.

$$PSE^{\pi}_{x_1,x_0}(y) = \mathbb{P}(y_{x_1|_\pi,x_0|_{\overline{\pi}}}) - \mathbb{P}(y_{x_0}) \tag{5.4}$$

where $\mathbb{P}(y_{x_1|_\pi,x_0|_{\overline{\pi}}})$ is the probability of $Y = y$ in the counterfactual situation where the effect of X on Y with the intervention $do(X = x_1)$ is transmitted along π, while the effect of X on Y without the intervention (x_0) is transmitted along paths not in π (denoted by: $\overline{\pi}$). For example, in the graph of Figure 5.2, if $\pi = X \rightarrow E \rightarrow Y$, then $\overline{\pi}$ includes $X \rightarrow Y$ and $X \rightarrow R \rightarrow Y$.

Other notions of causal fairness

In addition to total effect (TE) and related mediation analysis notions (NDE, NIE, and PSE), causal notions of fairness include *no unresolved discrimination* [110], *no proxy discrimination* [110] and *counterfactual fairness* [111].

No unresolved discrimination [110] is a fairness notion focusing on indirect causal effects from the sensitive variable X to the outcome Y. Unresolved discrimination is satisfied when no directed path from A to Y is allowed, except through a resolving (explaining) variable E. A resolving variable is any variable in a causal graph that is influenced by the sensitive attribute in a manner that is accepted as non-discriminatory.

Similarly to no unresolved discrimination, *no proxy discrimination* [110] focuses on indirect discrimination. A causal graph exhibits a potential proxy discrimination if there exists a path from the protected attribute X to the outcome Y that is blocked by a proxy/redlining variable R. It is called a proxy because it is used to decide about the outcome Y while it is a descendent of X that is significantly correlated with it in such a way that using the proxy in the decision has almost the same impact as using X directly. An outcome variable Y does not exhibit proxy discrimination if equality:

DEFINITION 5.3.5.

$$\mathbb{P}(Y \mid do(R = r)) = \mathbb{P}(Y \mid do(R = r')) \quad \forall\, r, r' \in dom(R) \tag{5.5}$$

holds for any potential proxy variable R.

The use of both no unresolved discrimination and no proxy discrimination in real scenarios is limited by the assumption of valid causal graph availability. Hence, both fairness notions depend on the correct output of the causal discovery task.

Counterfactual fairness [111] is a very strong fairness notion that requires equality between the observed outcome and the counterfactual outcome for every individual. That is, an outcome Y is counterfactually fair if under any assignment of values $\mathbf{V} = \mathbf{v}$ and any individual in the population,

DEFINITION 5.3.6.

$$\mathbb{P}(y_{x_1} \mid \mathbf{V} = \mathbf{v}, X = x_0) = \mathbb{P}(y_{x_0} \mid \mathbf{V} = \mathbf{v}, X = x_0) \tag{5.6}$$

where $\mathbf{V}$ represents the set of all remaining variables (all variables in the causal graph except $\{X, Y\}$). Counterfactual fairness, as an individual fairness notion, is satisfied if the probability distribution of the outcome Y is the same in the actual and counterfactual worlds, for every possible individual. For an exhaustive list of causal-based fairness notions, we refer interested readers to the survey of Makhlouf et al. [19].

5.4 Causal Assumptions

Causal inference is always subject to causal assumptions. Here we provide the main causal assumptions for potential outcome and Pearl's causal frameworks.

SUTVA [112] (Potential Outcome) The SUTVA (Stable Unit Treatment Value Assumption) requires that no influence on the treatment effect is induced by the interaction between individuals or the treatment mechanism [50].

Ignorability [112] (Potential Outcome) The Ignorability assumption requires that the sensitive attribute and the outcome are independent given observable variables.

Positivity [112] (Potential Outcome) The positivity is satisfied if all combinations of values of the treatment variable and covariates have nonzero probability.

Causal Graph [46] (DAG) Causal graph is the main requirement in the DAG framework and provides the complete specification of the relationships in the data. Formally, a causal graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$, composed of a set of variables/vertices $\mathbf{V}$ and a set of edges $\mathcal{E}$, is a directed acyclic graph (DAG) that describes the causal relations between variables. It is subject to further assumptions of the Causal Markov condition, Causal Faithfulness, and Causal Sufficiency. All three assumptions together encode the same requirements as the SUTVA and Ignorability in Potential Outcome.

Causal Markov condition [46] (DAG): Causal Markov condition is related to graph representation and requires every node to be independent of its non-descendants given its parents. Formally, a directed acyclic causal graph $\mathcal{G}$ with a set of vertices **V** and a probability distribution $\mathcal{P}$ over the vertices **V** satisfies the Causal Markov Condition if every node X in **V** is independent of $NonDescendants(X)$ given $Parents(X)$. It implies the absence of cycles, a requirement equivalent to the first assumption in SUTVA.

Causal faithfulness (DAG): Causal faithfulness is also a characteristic of a causal graph. It requires that all the conditional independence relations that hold in the data be encoded in the graph. Formally, a causal graph $\mathcal{G}$ and a probability distribution $\mathcal{P}$ on the same variables **V** are faithful to each other if all and only the conditional independence relations that hold in $\mathcal{P}$ are entailed by the Markov condition and the d separation in $\mathcal{G}$. For example, faithfulness would be violated if negatively correlated indirect paths in the data cancel out positively correlated directed paths rending Nationality and Visa decision variables independent, even though they are dependent (connected with directed edges) in the graph.

Causal sufficiency [46] (DAG): Causal sufficiency requires that there be no latent (hidden) confounders between variables in the data. It corresponds to the assumption of ignorability in the potential outcome framework.

5.5 Causal Discovery

Causal discovery is concerned with the identification of causal relationships from the data. More precisely, it aims to learn the fully directed DAG or partly directed PDAG that best describes the given data set. Causal discovery is performed by using observed samples. Previous knowledge of data structure can also be incorporated to facilitate the learning process. There are several causal discovery algorithms for a wide range of different assumptions; for a survey, see [113].

5.5.1 Causal discovery algorithms

The Peter and Clark (PC) [114] algorithm is a constraint-based method with two primary stages. The initial stage, known as "adjacency search," involves identifying the undirected skeleton of the Directed Acyclic Graph (DAG). The second stage focuses on estimating a completed partially directed acyclic graph (CPDAG). PC can be applied to linear, Gaussian data (the Fisher Z test), discrete multinomial data (the Chi-square test), and mixed multinomial/Gaussian data (the Conditional Gaussian test). PC uses an alpha parameter which is a cutoff, which signifies the threshold at which test results are considered indicative of dependence in a statistical test of independence and typically defaults to 0.05. When using a higher alpha value, it leads to a sparser graph. In other words, a higher alpha makes the test more stringent, and it requires stronger evidence to conclude that variables are dependent, resulting in fewer edges in the graphical model.

FCI (Fast Causal Inference) [115] algorithm is a constraint-based method designed to work with sample data, and it can also consider optional background knowledge. In the large sample limit, FCI provides an equivalence class of Conditional Bayesian Networks (CBNs) that encompass

the set of conditional independence relations believed to be valid in the population, even when there are hidden confounding variables. However, FCI has limitations and is most suitable for datasets with several thousand variables. When applied to realistic sample sizes, it can be inaccurate in determining both adjacencies and orientations.

FCI consists of two phases: the adjacency phase and the orientation phase. During the adjacency phase, the algorithm begins with a complete undirected graph and then performs a series of conditional independence tests. These tests lead to the removal of edges between pairs of variables that are determined to be independent given some subset of the observed variables. Conditioning sets that result in the removal of an edge are stored. By the end of the adjacency phase, the undirected graph correctly represents the set of adjacencies among variables, but all edges remain unoriented. FCI then proceeds to the orientation phase, where it uses the stored conditioning sets to orient as many edges as possible, adding directionality to the graph.

FGES [116] is an enhanced and parallelized variant of the Greedy Equivalence Search (GES) algorithm, initially developed by [117] and later studied by [118]. GES is a Bayesian algorithm that uses a heuristic approach to explore the space of Conditional Bayesian Networks (CBNs) and identify the model with the highest Bayesian score. Specifically, GES begins its search with an empty graph and proceeds with a forward-stepping search, where it adds edges between nodes to maximize the Bayesian score. This process continues until no further single-edge addition improves the score. Subsequently, it performs a backward-stepping search, eliminating edges until no single edge removal can enhance the score. These algorithms are capable of handling both continuous data, utilizing the Structural Equation Modeling Bayesian Information Criterion (SEM BIC) score, and discrete data, making use of the Bayesian Dirichlet equivalent uniform (BDeu) score. FGES takes the penalty discount parameter. A higher penalty discount value yields sparser graphs.

Iterative MCMC [119] is a hybrid optimization technique based on Markov chain Monte Carlo (MCMC) methods. The algorithm's initial step involves generating a skeleton, obtained through the Greedy Equivalence Search (GES) algorithm. Subsequently, it conducts a score-based search within the space defined by this initial skeleton, exploring various Directed Acyclic Graphs (DAGs).

The Max-min hill-climbing (MMHC) [120] is a hybrid approach that follows a two-step process. First, it estimates the skeleton of a Directed Acyclic Graph (DAG) using an algorithm known as Max-Min Parents and Children. Then, it applies a greedy hill-climbing search to determine the orientation of edges within the graph based on Bayesian scoring. MMHC is particularly suitable for domains with a high number of dimensions.

RECI by [121] (regression error-based causal inference) This algorithm addresses non-deterministic and non-linear relations and allows dependency between cause and noise. The algorithm's key idea is to fit regression models in both possible directions and to compare the MSE. Independence tests are not used, but the assumptions on the model depend on the regressor used for the model. In our experiments we used a polynomial regressor of degree 3 after rescaling to $[0, 1]$.

IGCI by [122] The Information Geometric Causal Inference model is a pairwise causal

discovery model that is able to determine the causal relationship in a deterministic setting $Y = f(X)$ (where f is invertible), under the 'independence assumption' $Cov[\log f', p_X] = 0$. In our experiments we have used a Gaussian reference measure and the sp1 or "1-spacing" method for entropy estimation used in [67].

ANM by [123] The approach assumes that $Y = f(X) + E$, where f is nonlinear. Causal inference is based on the independence between X and E. Parameters like Gaussian process regression and normalized HSIC can be used for the prediction and evaluation of the causal direction.

6 Other

The Expectation-Maximization Framework Let O be a random variable depending on an unknown parameter θ. Given that we observe $O = o$, the aim is to find the value of θ that maximizes the probability of this observation and, therefore, is *its best explanation*. To this purpose, we use the *log-likelihood function* $L(\theta) = \log P[O = o|\theta]$. A maximum likelihood estimate (MLE) of the parameter is then defined as $\operatorname{argmax}_\theta L(\theta)$ (which is the θ that maximizes $P[O = o|\theta]$, since log is monotone). The Expectation Maximization (EM) framework [124–126] is a powerful method to calculate $\operatorname{argmax}_\theta L(\theta)$.

6.1 Metrics for the quality of estimations

The Wasserstein distance This distance is defined between probability distributions in a metric space. Let $\mathcal{X}$ be a set provided with a distance d, and let μ, ν be two discrete probability distributions on $\mathcal{X}$. The Wasserstein distance between μ and ν is defined as

Definition 6.1.1.

$$\mathcal{W}(\mu, \nu) = \min_{\alpha} \sum_{x,y \in \mathcal{X}} \alpha(x, y)\, d(x, y), \tag{6.1}$$

where α represents a *coupling*, i.e., a joint distributions with marginals μ and ν satisfying the properties $\sum_{y \in \mathcal{X}} \alpha(x, y) = \mu(x)$ and $\sum_{x \in \mathcal{X}} \alpha(x, y) = \nu(y)$.

Accuracy Let X, Y be two random variables with support $\mathcal{X}$ and $\mathcal{Y}$ respectively, and joint distribution $\mathbb{P}[X, Y]$. Let $f : \mathcal{X} \to \mathcal{Y}$ be a function that, given $x \in \mathcal{X}$, *estimates* the corresponding y, and let $\hat{y}$ be the result, i.e. $\hat{y} = f(x)$. The accuracy of f is defined as the expected value of

$\mathbb{1}_{\hat{y}=y}$, the function that gives 1 if $\hat{y} = y$, and 0 otherwise. When the distribution is unknown, the accuracy is estimated empirically via a set of pairs $\{(x_i, y_i) \mid i \in \mathcal{I}\}$ independently sampled from $P[X, Y]$ (*testing set*), and is defined as:

DEFINITION 6.1.2.

$$Acc(\hat{Y}, Y) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}_{\hat{y}_i = y_i} \text{ where } \hat{y}_i = f(x_i). \tag{6.2}$$

Distortion If the variable E to be predicted ranges over a metric space, and the metric is important for decision-making, accuracy is not always the best way to measure the quality of the estimation. Arguably, it is more suitable to use the *distortion*, i.e., the expected distance between the true value and its estimation. The distortion in the estimation of E is defined as

DEFINITION 6.1.3.

$$Dist(\hat{E}, E) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} |\hat{e}_i - e_i|, \tag{6.3}$$

where $\hat{e}_i$ is an estimate of e_i.

Structural Hamming Distance (SHD) The Structural Hamming Distance is a metric used to measure the dissimilarity between two structural representations, often in the context of graph-based or hierarchical data. It is a way to quantify how different the structures of two entities, such as graphs or trees, are from each other. Unlike the traditional Hamming distance, which is used to measure the difference between binary strings, the Structural Hamming Distance considers the structural relationships and hierarchies within the data.

DEFINITION 6.1.4.

$$SHD(A, B) = \sum_{i=1}^{n} \delta(A_i, B_i) \tag{6.4}$$

Here A and B represent the two structural representations (e.g., graphs, trees) that are compared. A_i and B_i represent the structural elements or nodes within A and B, $\delta(A_i, B_i)$ is a function that measures the dissimilarity between the structural elements of A_i and B_i.

F1 Score The F1 Score is a popular metric used in binary classification to assess the model's accuracy and balance between precision [1] and recall [2]. It provides a single value that takes into account both false positives and false negatives. The F1 score is particularly useful when dealing with unbalanced data sets, where one class significantly outnumbers the other. The F1 Score is calculated using the following formula:

DEFINITION 6.1.5.

$$F1 = \frac{TP}{TP + 0.5(FP + FN)} \tag{6.5}$$

[1] Precision=$TP/(TP + FP)$
[2] Recall=$TP/(TP + FN)$

For the definitions of TP, FP, and FN see preliminaries on confusion matrix 3.1.

Part III

Fairness

7

BaBE: Enhancing Fairness via Estimation of Latent Explaining Variables

7.1 Introduction

One of the first notions of group fairness proposed in the literature was *statistical parity* (SP) 3.1, which enforces the probability of a positive prediction to be equal between different groups. However, SP has been criticized for causing a loss of accuracy and for ignoring circumstances that could justify the disparity. A more refined notion is *conditional statistical parity* (CSP) 3.3, which allows for some disparity as long as it is legitimated by explaining factors.

The most common pre-processing approach to achieve CSP (or an approximation of it) consists of editing the label Y (decision) in the training data, according to some heuristic, to ensure that the number of samples with $Y = 1$, $S = 1$, and $E = e$ is approximately the same number as those with $Y = 1$, $S = 0$, and $E = e$. However, one problem is that often E is not directly observable in the data, that is, it is a *latent* variable. Usually, we can observe some other variable Z that is representative of E, but the problem is that Z may be also influenced by the sensitive attribute S, hence Z itself can be biased. We illustrate this scenario with the following examples.

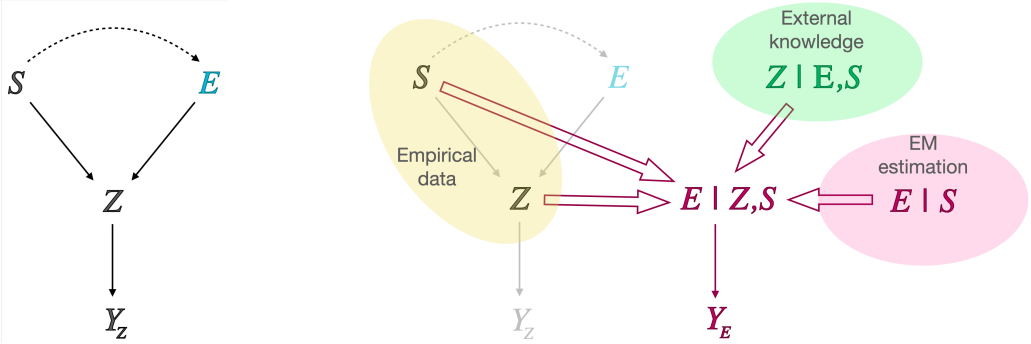


Figure 7.1: Left: illustration of the causal relation between the data. Right: illustration of our pre-processing method

Example 7.1.1. *The SAT (Scholastic Assessment Test) is a standardized test widely used for college admissions in the United States aiming at indicating the skill level of the applicant, and therefore her potential to succeed in college. However, the performance at the test can be affected by other socioeconomic, psychological, and cultural factors. For example, a recent study [127] indicates that, on average, black students are less likely to suffer the financial burden of retaking the test than white students. This causes a racial gap in scores, as retaking the test usually improves the result. Another study [128] reports that, on average, girls score approximately 30 points less on SAT than boys, despite the fact that girls routinely achieve higher grades in school. According to [128], the cause is the higher sensitivity to stress and test anxiety among females.*

Example 7.1.2. *Many healthcare systems in the United States rely on prediction algorithms to identify patients in need of assistance. One of the most used indicators is individual healthcare expenses, as they are easily available in insurance claim data. However, healthcare spending is influenced not only by health conditions, but also by socioeconomic status. A recent study [5] shows that the typical algorithms used by these healthcare systems are negatively biased against black patients, in the sense that, for the same prediction score, black patients are on average sicker than white ones. According to [5], this is due to bias in healthcare spending data, as black patients spend less on healthcare due to lower financial capabilities and lower level of trust in the white-dominated healthcare system and practitioners.*

Example 7.1.3. *Age is an important factor in determining retirement time, screening time for oncologic and cardiovascular diseases, and similar decisions. However, a chronological age (the time from birth to the current moment) does not always coincide with a biological age (an estimate based on biological markers that indicate the cellular age of an organism). Moreover, the error between biological age and chronological estimate is disparate for black and white populations. Recent studies show that black people on average are older biologically than chronologically, while white people have lower biological age compared to their chronological age [129–131].*

In the above examples, the "true skills", the "true health status" and the "true biological age", respectively, are the legitimate features E (explanation) on which we should base the decision. Unfortunately, E is not directly observable. What we can observe, instead, is the result of the SAT test and the healthcare-related spending, respectively. These are represented by the variable that we call Z. However, these indicators do not faithfully represent E, because they are also

influenced by other factors, namely economic status (or gender) and race, respectively. These are the sensitive attributes S.

The line of research that advocates the use of statistical parity [15, 132–135] adheres to the principle "we are all equal" [136], and makes the basic assumption that *E and S are independent*. However, in many cases, such as, for instance, in decisions regarding the medical treatment of genetic illnesses, race or gender could have a direct effect on the likelihood of the medical condition. For example, in our second running example, the real health status is on average lower in the black population due to socioeconomic factors. Hence, we allow the possibility of a link between the sensitive attribute S and the explaining value E, and aim to remove the discrimination introduced by the link between S and Z. The method we propose to remove the discrimination works equally well whether or not there is a link between S and E, and it does not modify this relation.

To summarize, in the original (unfair) scenario the decision Y is based on Z, which is influenced by both E and S. The situation is represented in Figure 7.1 (left). The arrow from S to Z represents that there is a causal relation between S and Z, and similarly for the other solid arrows[1], while the dashed arrow between S and E represents a relation that may or may not be present. To make a fair decision, we would like to base the decision Y only on E, but, as explained before, E may not be available directly. Therefore, we need to determine the most likely value of E for the given values of S and Z. To this purpose, we will derive the conditional distribution of E given Z and S, i.e. $\mathbb{P}[E|Z,S]$. The objective is illustrated in Figure 7.1 (right).

Note that E can be multidimensional and that we represent the effect of other possible latent variables by randomness in the distribution of the data. The method we propose uses a combination of the *Bayes theorem* and the *expectation-maximization method* (EM) [124], a powerful statistical technique for estimating latent variables as maximum likelihood parameters of empirical data observations. We call our method BaBE, for *Bayesian Bias Elimination*.

BaBE relies on some additional knowledge, namely an estimation of the conditional distribution of Z given S and E, that is, $\mathbb{P}[Z|E,S]$. This estimate can be obtained by collecting additional data. For example, for example 7.1.2, we could use the richer set of biomarkers, like in [5]. Alternatively, it can be produced by studies or experiments in a controlled environment. For example 7.1.1, we could assess the skills in some subjects by in-depth examinations and derive statistics on their SAT performance both at the first attempt and after a number of retakes.

One may question whether it is reasonable to assume that we can derive directly from the data $\mathbb{P}[Z|E,S]$ and not $\mathbb{P}[E|Z,S]$ (the derivation of the latter from the first is the essence of our proposal). We argue that, while it may not be *always* the case, there are real-life situations in which this assumption is justified, and in which, therefore, our method is applicable. Besides the above examples, one clear example is the study of symptoms (Z) induced by certain diseases (E): $\mathbb{P}[Z|E,S]$ can be statistically estimated from medical data collected by hospitals, while $\mathbb{P}[E|Z,S]$ cannot, because not all patients affected by symptoms necessarily enter the medical system. A further reason for assuming that we dispose of $\mathbb{P}[Z|E,S]$ and not of $\mathbb{P}[E|Z,S]$ is that the latter depends on the distribution of E, which can vary greatly depending on the geographical area, on

[1] Note that E is what in causality is called *a mediator*.

the social context, etc. For example, the racial health gap can be different in Europe and in the United States, where the experiments or data collection took place. In contrast, $\mathbb{P}[Z|E,S]$ may be more "universal", so it is convenient to invest in the estimation of the latter, which can be done once and then transferred to different contexts. Indeed, one advantage of our approach is that it allows for transferring of causal knowledge. That is, once we learn the relation $\mathbb{P}[Z|E,S]$, the method can be reused with a population with different proportions, that is, different $\mathbb{P}[E|S]$ (but the same $\mathbb{P}[Z|E,S]$). For more discussion on this point, we refer to [137–141].

Once $\mathbb{P}[E|Z,S]$ is estimated, we pre-process the training data by assigning a decision Y based on the most likely value e of E, for given values of S and Z. If e does not have enough probability mass, we may not achieve CSP, or even a good approximation of it. In this case, we can base the decision on a threshold for the estimated E, aiming instead at achieving equal opportunity (EO) [14], which we consider as a relaxation of the CSP. Formally, EO is classified as follows: $\mathbb{P}[\hat{Y}=1|Y=1,S=1] = \mathbb{P}[\hat{Y}=1|Y=1,S=0]$, where Y represents the "true decision", that is, the decision based on a threshold for the real value of E.

Related Work The notion of fairness that we consider in this work was introduced in [98] and it is known nowadays as *conditional statistical parity* (CSP) [142]. In [98], CSP is achieved through data prep-rocessing, by applying *local massaging* or *local preferential sampling* techniques. However, the authors consider only an *observable* explanatory variable E, not a *latent* E. Note that our Z, although observable, cannot be considered as an explanatory variable, because we are assuming it is influenced by the sensitive attribute in a way that would make it unfair to base the decision on Z. To better understand the difference, consider one of the main examples used in [98] to illustrate the idea, which is a kind of *Berkeley admission anomaly*, an instance of the *Simpson paradox* [143]. In this example, the admission to a certain university looks biased against women, but the disparity can actually be explained by the fact that female students tend to choose a more selective program. In this case, the explanatory variable is a mediator (the choice of the program) and is assumed to be legitimate as a cause of the disparity. In contrast, in our example, the observed score is considered to be influenced by social discrimination, hence it cannot be directly used as an explanatory variable.

The work closest to ours is [15], where there is a model containing a latent variable whose distribution is discovered using the expectation maximization method. However, in [15] the notion of fairness considered is *statistical parity* (SP). Using SP as a constraint (thus applying a sort of *self-fulfilling prophecy* approach) and other constraints such as the preservation of the total ratio of positive decisions, the authors determine what the distribution $\mathbb{P}[Z|E,S]$ should be, they distribute the probability mass uniformly on all attributes, and they finally apply the EM method to determine fair labels. In contrast, we are trying to find the most probable value of E for each combination of values of the other attributes (S and Z), to make a fair decision based on E, considered as the explanatory variable. We do not require statistical parity, nor do we assume a uniform distribution on all attributes. Instead, we use external knowledge as prior knowledge for applying the EM method. Another difference is that they optimize accuracy with respect to the observed biased labels, whereas we consider accuracy towards the true fair label dependent

on E, considered as the actual attribute on which the decision should be made.

Similar in spirit to [15], [134] tries to discover the latent variable that is maximally informative about the decision, while minimizing the correlation with the sensitive attribute (statistical disparity); this is done using a deep learning technique. Also, [16, 135, 144] use deep learning latent variable models. [16, 144] consider latent confounders and [135] considers the sensitive attribute as a confounder. The situations in which these assumptions apply are quite different from the problem we study since they aim at eliminating the effect of the confounder, while for us the latent variable is a mediator, and we want to use it as the basis for a fair decision. As a consequence, the notion of fairness that these works aim at achieving is not suitable for our case. [108] introduces path-specific counterfactual fairness, where (among other cases) they consider the latent cause of a mediator between the sensitive attribute and the decision. This is more similar to our notion of fairness. However, [108] assumes that the latent variable is independent of the sensitive attribute; as such, their method is not directly applicable to our problem. [132] uses probabilistic circuits to impose statistical parity and learn a relationship between the latent fair decision and other variables. Finally, [145] uses a notion of fairness called *disparate impact*, which is similar to statistical disparity, except that it is defined as a ratio (instead of a difference) between the probabilities of positive decisions for each group. Similarly to our work, [145] applies a corrective factor to the outcome of the observed variable Z, but its goal is to minimize the disparate impact (within a certain allowed threshold α), again in the spirit of minimizing statistical disparity. Their technique is also very different: they consider the distributions on the observed variable Z for each group, and they compute new distributions that minimize the earthmovers' distance and achieve the threshold α. Then they map each value of Z (for each group) in the new distribution to maintain the percentile.

7.2 Notation

$\hat{E}$, $\hat{Y}$ and Y notations In this chapter, $\hat{E}$ (with generic value $\hat{e}$) represents the estimation of the explanatory variable E. Similarly, $\hat{Y}_{\hat{E}}$ (with generic value $\hat{y}$) represents the decision estimation, based on $\hat{E}$, rather than the prediction of the model. To put it in context, recall that we propose a pre-processing method: $\hat{y}$ represents the value that we assign as decision in a sample of the training data during the pre-processing phase. Fairness and precision notions are defined with respect to these estimations. We use Y_Z to indicate the biased decision based on Z, and Y_E for the "true" decision based on E. When clear from the context, we may use Y instead of Y_E. We redefine SPD 3.1, CSPD 3.3 and EOD 3.2 using these notation:

Definition 7.2.1 (SPD).

$$\mathbb{P}[\hat{Y}_{\hat{E}} = 1|S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|S = 0] \tag{7.1}$$

Definition 7.2.2 (CSPD).

$$\mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 0] \tag{7.2}$$

DEFINITION 7.2.3 (EOD).

$$\mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 0] \tag{7.3}$$

7.3 The BaBE method

In this section, we describe the BaBE approach. We briefly recall the problem: we have a data model represented in Figure 7.1, where S is the sensitive attribute, E is the latent variable on which a fair decision should be based, and Z is an observed but biased version of E. We need to estimate the distribution $\mathbb{P}[E|Z, S]$. The first step is to estimate the distribution of E for each group, $\mathbb{P}[E|S]$. We accomplish this task by adapting the expectation maximization (EM) method to our particular setting. Then, from $\mathbb{P}[E|S]$ we derive, using the Bayes theorem, the estimation of $\hat{\mathbb{P}}[E|S, Z]$, from which we finally derive $\hat{E}$ and $\hat{Y}_{\hat{E}}$.

7.3.1 Derivation of BaBE as an instance of the EM method

In this section, we show how to apply the EM method to the problem we are considering, thus obtaining the main algorithm of our BaBE method.

Let E, Z and S be random variables on $\mathcal{E}$, $\mathcal{Z}$ and $\mathcal{S}$, with generic elements e, z and s respectively. Let $(\bar{z}, \bar{s}) = \{(z_i, s_i) \mid i = [1, \ldots, N]\}$ be a sequence of samples from the joint distribution $\mathbb{P}[Z, S]$, let

$$\bar{z}_s \stackrel{\text{def}}{=} \{z_i \ : \ i \in \{1, ..., N\} \wedge s_i = s\} \tag{7.4}$$

be the subsequence of $\bar{z}$ of elements paired with s in the samples and let M be $|\bar{z}_s|$. Then, the empirical probability of $Z = z$ given $S = s$ (i.e., the frequency of z in the samples with $S = s$) is defined as:

$$\varphi_s[z, \bar{z}_s] \stackrel{\text{def}}{=} \frac{|\{z_i \in \bar{z}_s \ : \ z_i = z\}|}{M}. \tag{7.5}$$

Now, given $(\bar{z}, \bar{s})$, $s \in \mathcal{S}$, $\varphi_s[z, \bar{z}_s]$ and the conditional distribution $\mathbb{P}[Z|E, S]$, we want to estimate the (unknown) $\mathbb{P}[E|S]$ by applying the expectation maximization (EM) method, that is, by finding the probability distribution on $\mathcal{E}$ that maximizes the probability of observing $\bar{z}_s$ given s (and therefore that is the best explanation of what we have observed). More precisely, we want to prove that our algorithm yields a Maximum Likelihood Estimation (MLE) $\hat{\mathbb{P}}[E|S]$ that approximates $\mathbb{P}[E|S]$. To this end, let Θ denote the set of all distributions on $\mathcal{E}$ conditioned on $S = s$, and let θ range over it. The *log-likelihood function* for $\bar{z}_s$ is $L_{\bar{z}_s} : \Theta \to \mathbf{R}$ such that

$$L_{\bar{z}_s}(\theta) \stackrel{\text{def}}{=} \log \mathbb{P}[\bar{Z}_s = \bar{z}_s|\theta] \tag{7.6}$$

where $\bar{Z}_s$ denotes a sequence of M random samples drawn from $\mathcal{Z}$ when $S = s$. Given $\bar{z}_s$, an unknown MLE $\mathbb{P}[E|S]$ is defined as $\text{argmax}_\theta \, L_{\bar{z}_s}(\theta)$, that is, as the θ that maximizes $L_{\bar{z}_s}(\theta)$ (and therefore $\mathbb{P}[\bar{Z}_s = \bar{z}_s|\theta]$, since log is monotone).

We now show how to adapt the EM framework to the above setting. We start by defining the

function

$$Q(\theta, \theta') \stackrel{\text{def}}{=} \mathbb{E}[\log \bar{\theta} \mid \bar{Z}_s = \bar{z}_s, S = s, \theta'] \tag{7.7}$$

where $\bar{\theta}$ denotes the probability distribution on sequences $\bar{e} = e_1, e_2, \ldots, e_M$ of i.i.d. events all with probability distribution θ. The above expectation is taken for all $\mathcal{E}$ and conditioned on $\bar{Z}_s = \bar{z}_s, S = s$, and assuming θ' as a prior approximation of $\mathbb{P}[E|S]$.

The function Q has the nice property $L_{\bar{z}_s}(\theta) - L_{\bar{z}_s}(\theta') \geq Q(\theta, \theta') - Q(\theta', \theta')$. Therefore, to improve the approximation of the MLE, that is, to find an estimation θ that improves the estimation θ', it is sufficient to compute $Q(\theta, \theta')$ and find the θ that maximizes it.

LEMMA 7.1.

$$Q(\theta, \theta') = \sum_{i=1}^{M} \sum_{e \in \mathcal{E}} \frac{\mathbb{P}[Z_s = z_i | E = e, S = s]\ \theta'[e|s]}{\sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z_i | E = e', S = s]\ \theta'[e'|s]} \log \theta[e|s].$$

Proof. Given that the E_is are i.i.d., by definition and linearity of conditional expectation, we have that:

$$\begin{aligned}
&\mathbb{E}[\log \bar{\theta} \mid \bar{Z}_s = \bar{z}_s, S = s, \theta'] \\
&= \mathbb{E}\left[\log \prod_{i=1}^{M} \theta[e_i|s] \;\middle|\; \bar{Z}_s = \bar{z}_s, S = s, \theta'\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{M} \log \theta[e_i|s] \;\middle|\; \bar{Z}_s = \bar{z}_s, S = s, \theta'\right] \\
&= \sum_{i=1}^{M} \mathbb{E}[\log \theta[e_i|s] \mid \bar{Z}_s = \bar{z}_s, S = s, \theta'] \\
&= \sum_{i=1}^{M} \sum_{e \in \mathcal{E}} \mathbb{P}[E = e | Z_s = z_i, S = s]\ \log \theta[e|s]
\end{aligned} \tag{7.8}$$

where $\mathbb{P}[E|Z_s, S]$ is a probability based on the estimation θ' of the unknown $\mathbb{P}[E|S]$. By taking the marginal distribution, we have that:

$$\begin{aligned}
&\mathbb{P}[Z_s = z_i | S = s] \\
&= \sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z_i, E = e' | S = s] \\
&= \sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z_i | E = e', S = s] \theta'[e'|s].
\end{aligned} \tag{7.9}$$

By the conditional Bayes theorem and (7.9), we have that

$$\begin{aligned}
&\mathbb{P}[E = e|Z_s = z_i, S = s] \\
&= \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\theta'[e|s]}{\mathbb{P}[Z_s = z_i|S = s]} \\
&= \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\theta'[e|s]}{\sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z_i|E = e', S = s]\theta'[e'|s]}
\end{aligned}$$

By plugging the latter equality into (7.8), we conclude the proof. □

The next Lemma tells us that $\hat{\mathbb{P}}[E|S]^{(t+1)}$ (as defined in Algorithm 1) is the distribution that maximizes $Q\left(\cdot, \hat{\mathbb{P}}[E|S]^{(t)}\right)$. This fact will allow us to conclude that the algorithm approximates the MLE $\text{argmax}_\theta L_{\bar{z}_s}(\theta)$.

Lemma 7.2. The θ that maximizes $Q(\cdot, \theta')$ is such that, for every $e \in \mathcal{E}$:

$$\theta[e|s] = \sum_{z \in \mathcal{Z}} \varphi_s[z, \bar{z}_s] \frac{\mathbb{P}[Z_s = z|E = e, S = s]\ \theta'[e|s]}{\sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z|E = e', S = s]\ \theta'[e'|s]}.$$

Proof. By the method of Lagrangian multipliers, we can find the θ that maximizes $Q(\theta, \theta')$ by adding to the latter the term $\lambda\left(\sum_{e \in \mathcal{E}} \theta[e|s] - 1\right)$, for some λ, and study the function

$$F(\theta, \theta') \triangleq Q(\theta, \theta') + \lambda\left(\sum_{e \in \mathcal{E}} \theta[e|s] - 1\right) \tag{7.10}$$

that has the same stationary points as $Q(\theta, \theta')$ since $\sum_{e \in \mathcal{E}} \theta[e|s] = 1$, being θ a probability distribution on $\mathcal{E}$ given $S = s$. To find the stationary points of F, we impose that all its partial derivatives, including the one w.r.t. λ, are equal to 0. For the latter one, we require that

$$\frac{\partial F}{\partial \lambda} = \sum_{e \in \mathcal{E}} \theta[e|s] - 1 = 0 \tag{7.11}$$

and this trivially holds since $\theta[\cdot|s]$ is a distribution for every s. For the former ones, by relying on Lemma 7.1, we impose that, for every $e \in \mathcal{E}$,

$$\begin{aligned}
&\frac{\partial F}{\partial \theta[e|s]} \\
&= \frac{1}{\theta[e|s]} \sum_{i=1}^{M} \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\ \theta'[e|s]}{\sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z_i|E = e', S = s]\ \theta'[e'|s]} + \lambda \\
&= 0
\end{aligned} \tag{7.12}$$

By multiplying the last equality by $\theta[e|s]$, we get:

$$\lambda\, \theta[e|s] = -\sum_{i=1}^{M} \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\ \theta'[e|s]}{\sum_{e' \in \mathcal{E}} \mathbb{P}[Z_s = z_i|E = e', S = s]\ \theta'[e'|s]}. \tag{7.13}$$

By summing both sides of (7.13) on all $e \in \mathcal{E}$, we obtain:

$$\begin{aligned}
&\lambda \sum_{e\in\mathcal{E}} \theta[e|s] \\
&= -\sum_{e\in\mathcal{E}} \sum_{i=1}^{M} \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\ \theta'[e|s]}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z_s = z_i|E = e', S = s]\ \theta'[e'|s]} \\
&= -\sum_{e\in\mathcal{E}} \sum_{z\in\mathcal{Z}} \varphi_s[z, \bar{z}_s] M \frac{\mathbb{P}[Z_s = z|E = e, S = s]\ \theta'[e|s]}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z_s = z|E = e', S = s]\ \theta'[e'|s]} \\
&= -M \sum_{z\in\mathcal{Z}} \varphi_s[z, \bar{z}_s] \frac{\sum_{e\in\mathcal{E}} \mathbb{P}[Z_s = z|E = e, S = s]\ \theta'[e|s]}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z_s = z|E = e', S = s]\ \theta'[e'|s]} \\
&= -M \sum_{z\in\mathcal{Z}} \varphi_s[z, \bar{z}_s] \\
&= -M
\end{aligned} \tag{7.14}$$

where the last step holds because of (7.4), and the second step holds because, again by (7.4), we have that, for any function f:

$$\sum_{i=1}^{M} f(z_i) = \sum_{z\in\mathcal{Z}} \varphi_s[z, \bar{z}_s] M f(z). \tag{7.15}$$

Hence, since $\theta[\cdot|s]$ is a probability distribution on $\mathcal{E}$, we obtain that (7.12) is satisfied by taking $\lambda = -M$.

Therefore, by isolating $\theta[e|s]$ from (7.13) and by using (7.15), we can conclude that, for every $e \in \mathcal{E}$, we have that

$$\begin{aligned}
&\theta[e|s] \\
&= -\frac{1}{\lambda} \sum_{i=1}^{M} \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\ \theta'[e|s]}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z_s = z_i|E = e', S = s]\ \theta'[e'|s]} \\
&= \frac{1}{M} \sum_{i=1}^{M} \frac{\mathbb{P}[Z_s = z_i|E = e, S = s]\ \theta'[e|s]}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z_s = z_i|E = e', S = s]\ \theta'[e'|s]} \\
&= \sum_{z\in\mathcal{Z}} \varphi_s[z, \bar{z}_s] \frac{\mathbb{P}[Z_s = z|E = e, S = s]\ \theta'[e|s]}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z_s = z|E = e', S = s]\ \theta'[e'|s]}
\end{aligned}$$

□

Now, for the given $s \in \mathcal{S}$, we define the sequence $\left\{ \hat{\mathbb{P}}[E|S = s]^{(t)} \right\}_{t\geq 0}$ as follows:

$$\hat{\mathbb{P}}[E|S = s]^{(0)} \stackrel{\text{def}}{=} \text{any fully supported distribution}$$

$$\hat{\mathbb{P}}[E|S = s]^{(t+1)} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} Q(\theta, \hat{\mathbb{P}}[E|S = s]^{(t)})$$

The next theorem states the key property of our algorithm, i.e. that $\left\{ \hat{\mathbb{P}}[E|S = s]^{(t)} \right\}_{t\geq 0}$

tends to the MLE $\text{argmax}_\theta L_{\bar{z}_s}(\theta)$. The proof of the theorem follows from the fact that $Q(\theta, \theta')$ has continuous derivatives in both its arguments, and from Theorem 4.3 in [146] (which is a reformulation of a result due to Wu [126]).

THEOREM 7.3. $\lim_{t\to\infty} \hat{\mathbb{P}}[E|S=s]^{(t)} = \operatorname{argmax}_\theta L_{\bar{z}_s}(\theta)$.

Furthermore, if $\mathbb{P}[Z|E,S]$, seen as a stochastic matrix, is invertible, then the MLE $\operatorname{argmax}_\theta L_{\bar{z}_s}(\theta)$ is unique. The proof follows from Theorem 4 in [147].

7.3.2 Deriving $\hat{\mathbb{P}}[E|S]$

We estimate the unknown parameter $\mathbb{P}[E|S]$ as the MLE of a sequence of samples $(\bar{z}, \bar{s}) = \{(z_i, s_i) \mid i \in [1, N]\}$,[2] assuming that we know the effect of the bias, i.e., $\mathbb{P}[Z|E,S]$.

We denote by $\varphi_s[z, \bar{z}]$ the empirical probability of $Z = z$ given $S = s$, i.e., the frequency of z in the samples with $S = s$. Algorithm 1 estimates $\mathbb{P}[E|S]$ by starting with the uniform distribution and by iteratively computing at step t a new estimation $\hat{\mathbb{P}}[E|S]^{(t)}$ from the previous one, getting closer and closer to the MLE. In the additional material we show how Algorithm 1 is obtained from the EM method.

Algorithm 1 : Estimation of $\mathbb{P}[E|S]$

Input Data: $\{(z_i, s_i) | i \in [1, \dots N]\}$, $\mathbb{P}[Z = z|E = e, S = s]\}$ and γ (desired level of precision)
Result: An approximation (up to γ) $\hat{\mathbb{P}}[E|S]$ of the MLE

Compute $\varphi_s[z, \bar{z}]$ for all z, s
$\hat{\mathbb{P}}[E = e|S = s]^{(0)} = \frac{1}{|E|}$ for all e, where $|E|$ is the cardinality of the domain of E.
$t = 0$
repeat
  $t = t + 1$
  $\hat{\mathbb{P}}[E = e|S = s]^{(t)} = \sum_{z\in\mathcal{Z}} \varphi_s[z, \bar{z}] \frac{\mathbb{P}[Z=z|E=e,S=s]\hat{\mathbb{P}}[E=e|S=s]^{(t-1)}}{\sum_{e'\in\mathcal{E}} \mathbb{P}[Z=z|E=e',S=s]\hat{\mathbb{P}}[E=e'|S=s]^{(t-1)}}$, for all e, s
until $\left|\hat{\mathbb{P}}[E = e|S = s]^{(t)} - \hat{\mathbb{P}}[E = e|S = s]^{(t-1)}\right| < \gamma$, for all e, s
return $\hat{\mathbb{P}}[E|S] = \hat{\mathbb{P}}[E|S]^{(t)}$

7.3.3 Deriving $\hat{\mathbb{P}}[E|Z,S]$ from $\hat{\mathbb{P}}[E|S]$

Given the data $\{(z_i, s_i) \mid i \in [1, N]\}$, the conditional distributions $\mathbb{P}[Z|E,S]$, and the estimation $\hat{\mathbb{P}}[E|S]$, we use the Bayes formula to estimate $\hat{\mathbb{P}}[E = e|Z = z, S = s]$ as

$$\hat{\mathbb{P}}[E = e|Z = z, S = s] = \frac{\mathbb{P}[Z=z|E=e,S=s]\hat{\mathbb{P}}[E=e|S=s]}{\mathbb{P}[Z=z|S=s]}$$

7.3.4 Deriving $\hat{E}$ and $\hat{Y}_{\hat{E}}$ from $\hat{\mathbb{P}}[E|Z,S]$

We propose two ways to derive $\hat{Y}_{\hat{E}}$ for pre-processing the samples in the training data, depending on how much probability mass is concentrated on the mode of $\hat{\mathbb{P}}[E|Z,S]$. We denote by τ the threshold for the values of E that qualify for the positive decision.

[2] We use the notation $[a, b]$ to represent the integers from a to b.

Method 1 Given z and s, if $\hat{\mathbb{P}}[E|Z=z,S=s]$ is unimodal and has a large probability mass (say, 50% or more) in its mode, then we can safely set $\hat{E}$ as that mode. Namely, if $\max_e \hat{\mathbb{P}}[E=e|Z=z,S=s] \geq 0.5$ then we set $\hat{e} = \text{argmax}_e \hat{\mathbb{P}}[E=e|Z=z,S=s]$, and we can then use $\hat{e}$ directly to set $\hat{Y}_{\hat{E}}=1$ or $\hat{Y}_{\hat{E}}=0$ in those samples with $Z=z$ and $S=s$, depending on whether $\hat{e} \geq \tau$ or not, respectively. Our experimental results show that this method gives good accuracy.

Method 2 If $\hat{\mathbb{P}}[E|Z=z,S=s]$ is dispersed in several values, so that no value is strongly predominant, then it is impossible to estimate individual values for E with high accuracy. However, we can still accurately estimate Y_E as follows: Let $\sigma_0 = \sum_{e<\tau} \hat{\mathbb{P}}[E=e|Z=z,S=s]$ and $\sigma_1 = \sum_{e\geq\tau} \hat{\mathbb{P}}[E=e|Z=z,S=s]$. If $\sigma_0 < \sigma_1$, then we set $\hat{Y}_{\hat{E}}=1$; otherwise, $\hat{Y}_{\hat{E}}=0$.

7.4 Experiments

In this section, we test BaBE on scenarios corresponding to Examples 7.1.1 and 7.1.2, using synthetic data sets and a real data set respectively, and we compare our results with those achieved by the following well-known pre-processing approaches that aim to satisfy statistical parity.

Metrics

We will use the following metrics to measure fairness: Statistical parity difference (SPD) 7.1, Conditional statistical parity difference ($CSPD_e$) 7.2, Equal opportunity difference (EOD) 7.3. The performance is measured by accuracy ($Acc(\hat{Y},Y)$) 6.2, distortion ($Dist(\hat{E},E)$) 6.3, and the Wasserstein distance between the true and estimated distributions ($\mathcal{W}(\mu,\nu)$) 6.1.

Other Algorithms for Comparison

The first approach we compare with is the *disparate impact* (DI) *remover* [145, 148] [3]. DI has a parameter λ, which represents the minimum allowed ratio between the probability of success ($\hat{Y}=1$) of each group (hence $\lambda=1$ corresponds to statistical parity). For the experiments, we use $\lambda=0.8$.

The second algorithm we compare with ours is the *naive Bayes* (NB) [15] [4]. NB also applies the EM method; however, in contrast to our work, NB assumes that E and S are independent, and uses EM to take decisions that optimize the trade-off between SPD and accuracy.

Synthetic data sets with varying means for $E|S=1$ and $E|S=0$

We generate synthetic data sets as follows. First, we generate a data set (multiset) of 20K elements $\{s_i\}_{i\in[1,20\text{K}]}$ representing values for the sensitive variable (group) S, where each s_i is sampled from the Bernoulli distribution $\mathcal{B}(0.5)$. This means that the two groups are about even. Then, we set the domain of E to be equal to $[0,99]$, and to each of the elements s_i in

[3] We use the implementation by [148].
[4] Implementation kindly provided by the authors of [15].

the sequence we associate a value e_i for the variable E, sampled from the normal distribution $\mathcal{N}(mean0, sd)$ if $s_i = 0$ and from $\mathcal{N}(mean1, sd)$ if $s_i = 1$,[5] where the mean *mean1* is set to 60, and the standard deviation sd is set to 30. On the other hand, the value of *mean0*, varies through the experiments from 40 to 80. Varying *mean0* will allow us to test how our method behaves when E is independent of S or not. Finally, to each pair (s_i, e_i) we associate a value z_i with Z by applying a bias to e_i with a certain probability. More precisely, $z_i = e_i + (bias \times e_i)$, where *bias* is sampled from $\mathcal{N}(-0.2, 0.05)$ if $s_i = 0$ and from $\mathcal{N}(0.2, 0.05)$ if $s_i = 1$. The threshold for the decision is $E = 60$, namely: $Y = 1$ if $E > 60$ and $Y = 0$ otherwise.

Synthetic data sets with distributions shifts in $E|S = 1$ and $E|S = 0$

This group of data sets is generated to test the transfer of causal knowledge to populations with different distributions. For this purpose, we estimate the distributions $\mathbb{P}[Z|E, S]$ from synthetic data generated as in the first group of experiments, with *mean0* = 60. We call these "source data". Then we generate different populations where *mean0* varies from 40 to 80. The percentage of the two groups in these new populations also changes: we have set the group 1 to be 60% of the population, and, consequently, the group 0 to be 40%.

The real-world data set

The National Health and Nutrition Examination Survey (NHANES) is a series of studies that are intended to evaluate the health and nutritional status of adults and children in the United States. The survey is unique in that it incorporates in-depth interviews and detailed physical examinations. Health-related questions and demographics are included in the NHANES interview. For the survey, the sample was selected to represent the US population of all ages. To produce reliable statistics, NHANES oversamples individuals aged 60 and over, African Americans, and Hispanics. The National Health and Nutrition Examination Survey (NHANES) [6] conducted by the National Center for Health Statistics is a popular source for studying biological aging [149–152]. The data set consists of 8243 samples. For our experiments, we use three variables from the data set, race (black or white), which is out S, chronological age (20-90), which is our Z, and an estimate of the biological age of the original KDM [7] biological age (variable 'kdm0') which is our E. We choose chronological or biological age 75 or more as the threshold to set $Y_Z = 1$ and $Y_E = 1$. This age group shows the most racial disparity in biological aging in the NHANES data set (Figure 7.2). Additionally, it is a reasonable age to check for age-related diseases or consider retirement.

[5] To keep the samples in the range of E, we re-sample the values that are lower than 0 or higher than 99. We also discretize them by rounding to the nearest integer.

[6] https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

[7] Klemera and Doubal's method for calculating the biological age from the set of biomarkers.

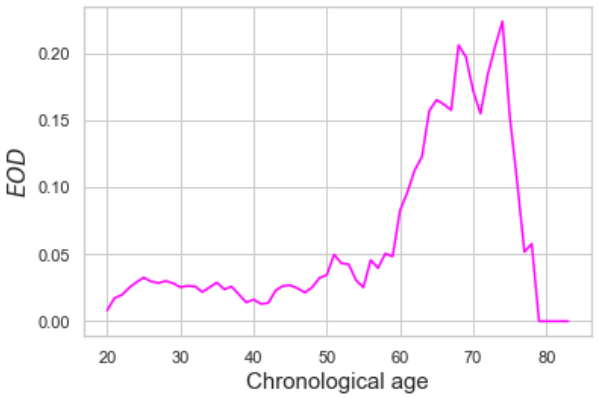


Figure 7.2: The graph shows the equal opportunity difference (EOD) between the Y_E and Y_Z, when different thresholds for Z (chronological age) are selected. The disparity is largest around the chronological age equal 75 years.

Application of BaBE

Once the data are generated, we use a random portion of them (50%) to derive $\mathbb{P}[Z|E,S]$ and $\mathbb{P}[E|S]$, which we consider as the "true" distributions. Then we take another portion of the data (40%) randomly selected from the unused ones, remove the E values from them, and use them to compute the empirical distribution $\mathbb{P}[Z|S]$ and to produce, by applying our BaBE method, the estimates $\hat{\mathbb{P}}[E|S]$ and $\hat{\mathbb{P}}[E|Z,S]$. We verify that these satisfy the conditions for Method 1, and we apply this method to set the values of $\hat{E}$ and $\hat{Y}_{\hat{E}}$ for each sample. In the second set of experiments, "source data" is not available for BaBE. The prior knowledge in the form of $\mathbb{P}[Z|E,S]$ derived from the "source data" is applied to estimate $\mathbb{P}[E|S]$ in the data sets, where it is different from the "source data". Experiments on the NHANES data are carried out using Method 2 of the BaBE application. We evaluated various metrics for the precision of the estimations and fairness and compared the performance of BaBE with disparate impact remover (DI) and with naive Bayes (NB). The boxplots are obtained by repeating the experiments ten times with the same parameters. We report the results for the values of $mean0$ equal to 40,60 and 80.

Experimental results on the synthetic data with varying means for $E|S=1$ and $E|S=0$

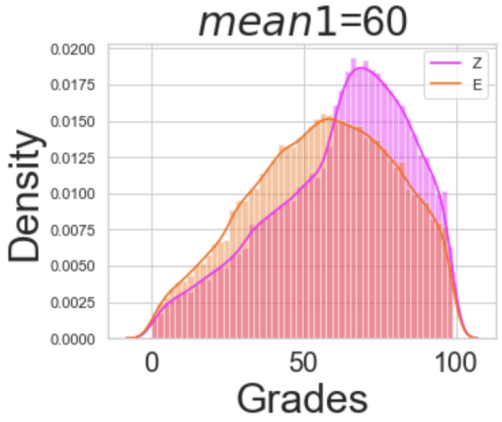


Figure 7.3: The distributions $\mathbb{P}[E|S=1]$ (orange) and $\mathbb{P}[Z|S=1]$ (magenta)

The distributions of E and Z in the data set, for each group $S = 1$ and $S = 0$, are shown in Figures 7.3 and 7.4 respectively.

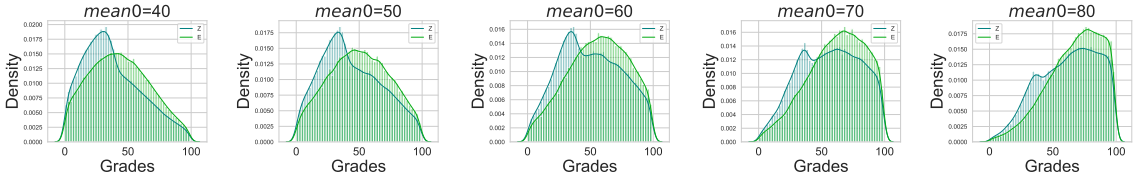


Figure 7.4: The distributions of E (green) and Z (blue) for $S = 0$, i.e., $\mathbb{P}[E|S = 0]$ and $\mathbb{P}[Z|S = 0]$, respectively

We now apply our BaBE method to estimate the distributions $\mathbb{P}[E|S = 1]$ and $\mathbb{P}[E|S = 0]$. The corresponding estimates $\hat{\mathbb{P}}[E|S = 1]$ and $\hat{\mathbb{P}}[E|S = 0]$ are shown in Figures 7.5 and 7.7, respectively. As we can see, all the estimates are very close to the original distributions. We note that the estimation tends to exaggerate the irregularities in the original distribution. This is a characteristic of the EM method when the volume of data is not very large.

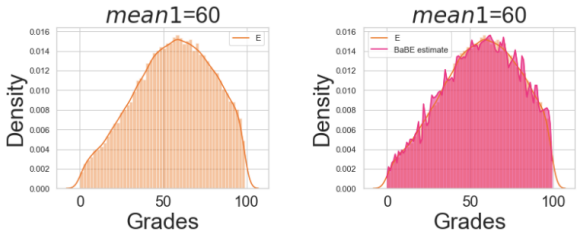


Figure 7.5: The original distribution $\mathbb{P}[E|S = 1]$ (orange), and the estimate $\hat{\mathbb{P}}[E|S = 1]$ produced by BaBE (magenta)

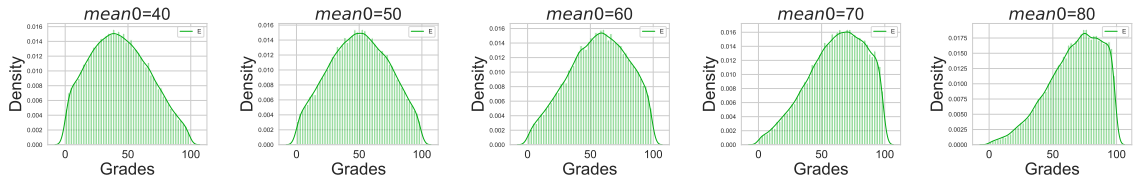


Figure 7.6: The original distributions $\mathbb{P}[E|S = 0]$.

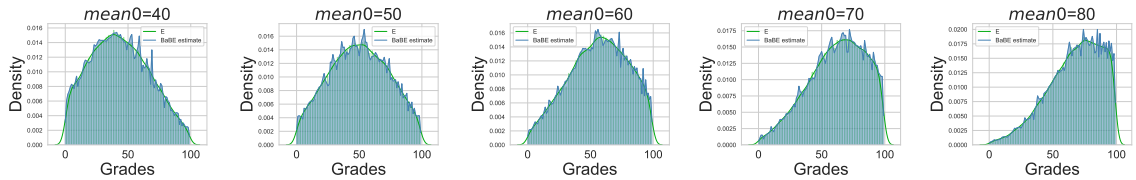


Figure 7.7: The original distributions $\mathbb{P}[E|S = 0]$ (green), and the estimates $\hat{\mathbb{P}}[E|S = 0]$ produced by BaBE (blue)

We now apply the DI method to estimate the E values in the data sets (obtained by applying a correction to Z), and from the resulting data sets we compute (by counting the frequencies) the distributions of the modified E for each group. The corresponding distributions are shown in Figures 7.8 and 7.9, respectively. Note that the new distributions are not very close to the originals, but, on the other hand, estimating the true E is not the goal of DI. Rather, DI aims at making the distributions of E for the two groups as similar as possible, thus reducing the statistical parity difference. DI achieves the goal by applying a negative correction on Z for group 0, and positive for group 1. For this reason, even though there is only one data set for group 1, we get 5 different distributions of E, one for each value of *mean0*.

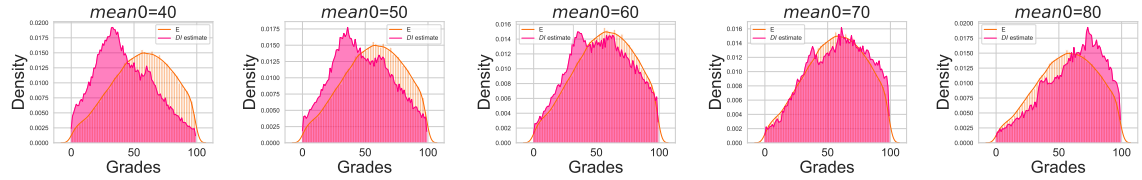


Figure 7.8: The original distribution $\mathbb{P}[E|S=1]$ (orange), and the distributions of the E estimated by DI (magenta) for group 1

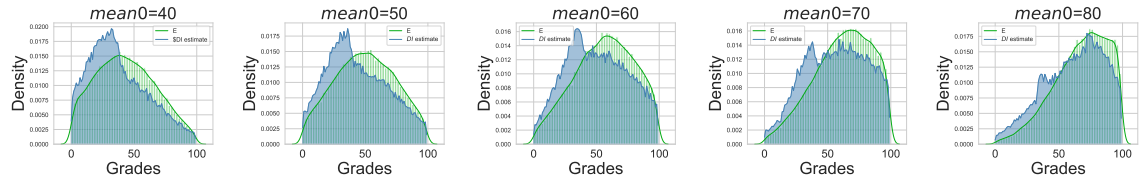


Figure 7.9: The original distributions $\mathbb{P}[E|S=0]$ (green), and the distributions of the E estimated by DI (blue) for group 0

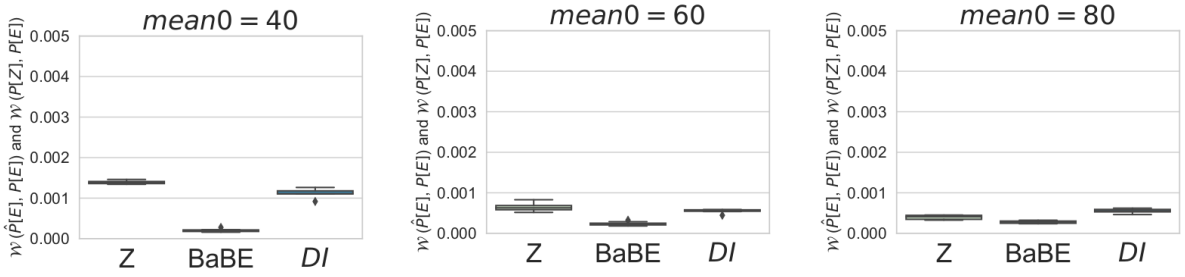


Figure 7.10: Wasserstein distance.

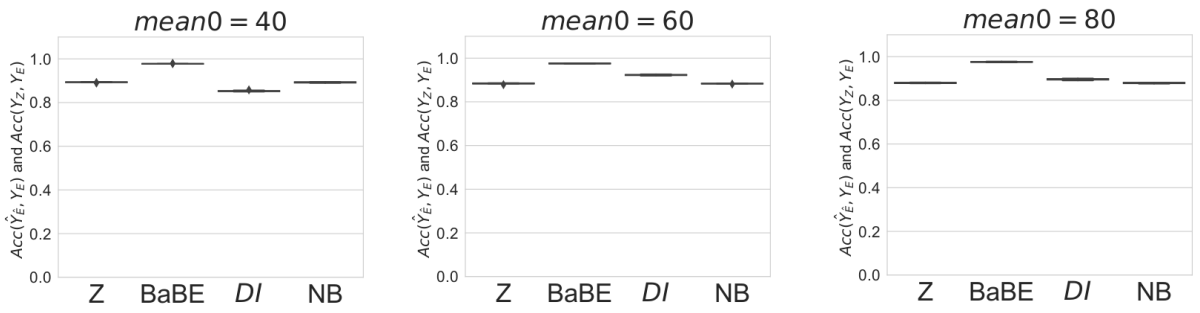


Figure 7.11: Accuracy.

Figure 7.10 shows the Wasserstein distance (6.1) between $\mathbb{P}[Z]$ and $\mathbb{P}[E]$ and between the estimate $\hat{\mathbb{P}}[E]$ and $\mathbb{P}[E]$ for BaBE and DI (NB does not estimate E). As expected, BaBE's estimate is quite accurate.

Figure A.6 shows the accuracy (6.2) of the prediction Y_Z based on Z and the accuracy of the prediction $\hat{Y}_{\hat{E}}$ obtained by BaBE, DI, and NB, respectively. As expected, BaBE's method produces more accurate predictions, since we aim at achieving CSP rather than SP.

Figure 7.12 shows the distortion (6.3) of Z and of $\hat{E}$ for BaBE and DI (NB does not estimate E). BaBE's results are quite good: the average difference is only a couple of grades.

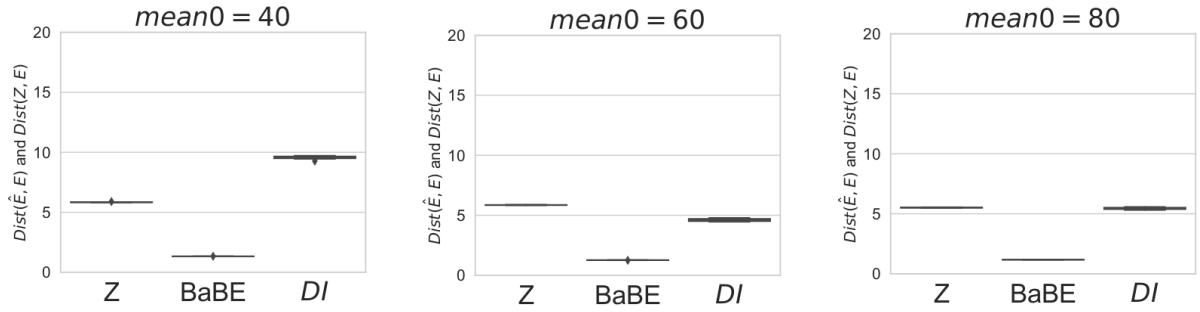


Figure 7.12: Distortion.

Figure 7.13 compares the statistical parity difference (SPD) of the prediction $\hat{Y}_{\hat{E}}$ obtained with various methods, and the SPD of Y_Z. The SPD for $\hat{Y}_{\hat{E}}$ is defined in (7.1), for Y_Z is defined as $\mathbb{P}[Y_Z = 1|S = 1] - \mathbb{P}[Y_Z = 1|S = 0]$. Note that the SPD of Y_Z decreases as *mean0* (the merit of group 0) increases. In particular, it becomes very small when *mean0* = 80. This is because in this case the merit of the group 0 is greater than that of the group 1 (we recall that *mean1* = 60), which compensates the effect of the bias (negative for group 0 and positive for group 1). Regarding $\hat{Y}_{\hat{E}}$, we recall that DI and NB are designed to optimize SPD under some constraints, while BaBE is not. As expected, with BaBE the SPD is quite large in all cases except when the merit is similar for the two groups (*mean0* = *mean1* = 60). DI has quite good results (very low SPD) in all cases. In contrast, NB performs poorly. We think this is because the parameters of this experiment clash with the constraints of NB. In other experiments, NB performs well (cf. supplementary material).

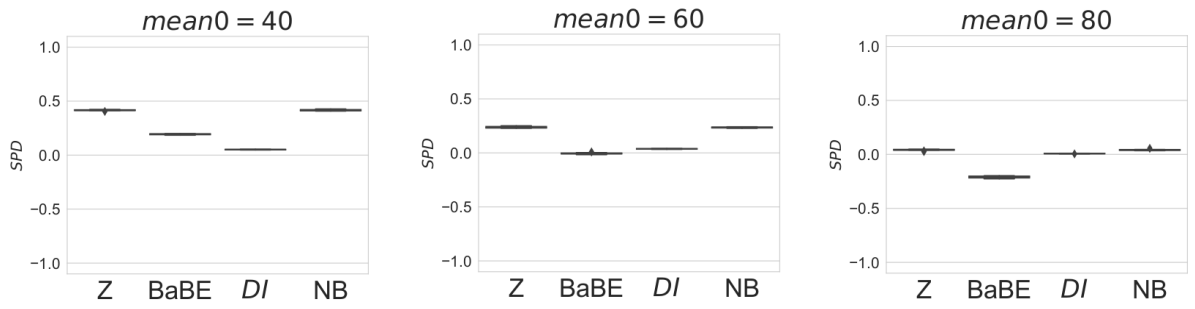


Figure 7.13: Statistical Parity Difference (SPD).

Figure 7.14 compares the Conditional Statistical Parity Difference (CSPD) of the $\hat{Y}_{\hat{E}}$ obtained with various methods and the CSPD of Y_Z. The CSPD for $\hat{Y}_{\hat{E}}$ is defined in (7.2), for Y_Z is defined as $\mathbb{P}[Y_Z = 1|E, S = 1] - \mathbb{P}[Y_Z = 1|E, S = 0]$. Note that BaBE performs very well in all cases, as expected. Also, DI performs surprisingly very well in the first two cases, but this is just a coincidence, as for *mean1* = 80 and in other experiments (cf. supplementary material) CSPD is high.

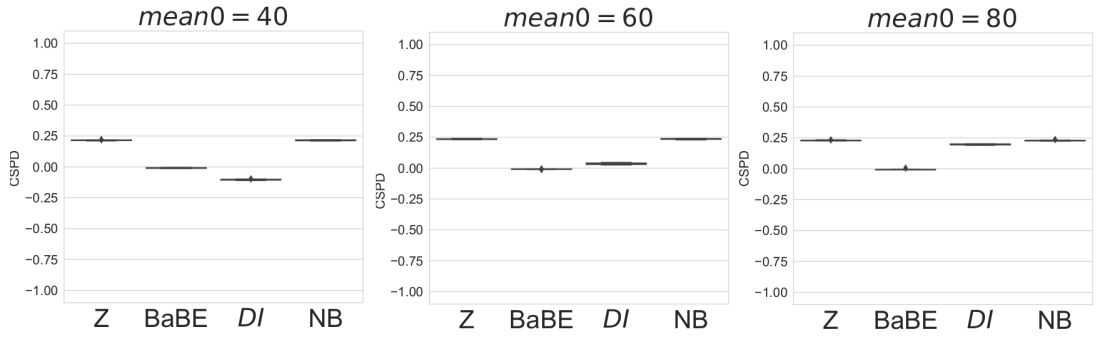


Figure 7.14: Conditional Statistical Parity Difference (CSPD).

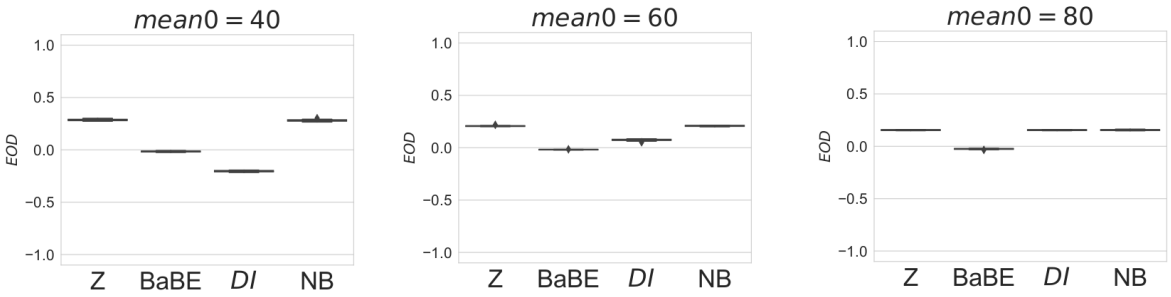


Figure 7.15: Equal Opportunity Difference.

Figure 7.15 compares the Equal Opportunity Difference (EOD) of the $\hat{Y}_{\hat{E}}$ obtained with the various methods and the EOD of Y_Z. The EOD for $\hat{Y}_{\hat{E}}$ is defined in (7.3), for Y_Z is defined as $\mathbb{P}[Y_Z = 1|Y_E = 1, S = 1] - \mathbb{P}[Y_Z = 1|Y_E = 1, S = 0]$. Again, BaBE performs very well in all cases, as expected. Note that in the first case DI does not perform well for EOD, in contrast to its performance for CSPD_e with $e = 55$. This is because EOD is computed for all e, and for other values of e the CSPD_e of DI is large.

Results on synthetic data sets with distribution shift

In this group of experiments, we show that BaBE is compatible with the transfer of causal knowledge to populations with different distributions. The distribution for the source data are shown in Figure 7.16, and those for the the new populations are shown in 7.17.

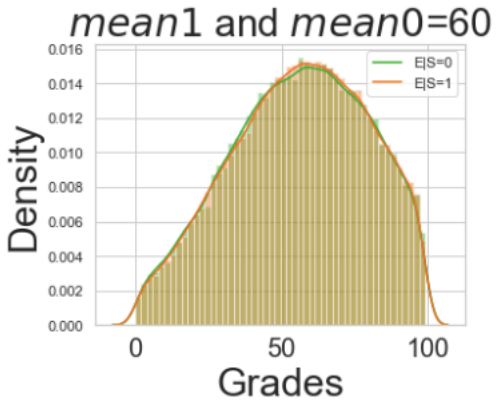


Figure 7.16: The distribution of $E|S$ in the source data for $\mathbb{P}[Z|E,S]$

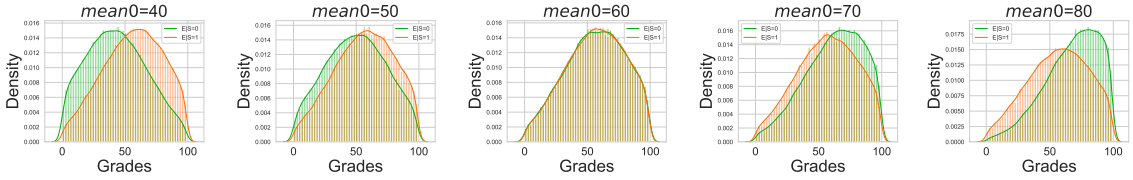


Figure 7.17: The distribution of $E|S$ in the new populations

Figure 7.18 shows the Wasserstein distances between the true distributions and the estimated ones. As we can see, BaBE manages to estimate E quite well: the distance w.r.t. E is very small.

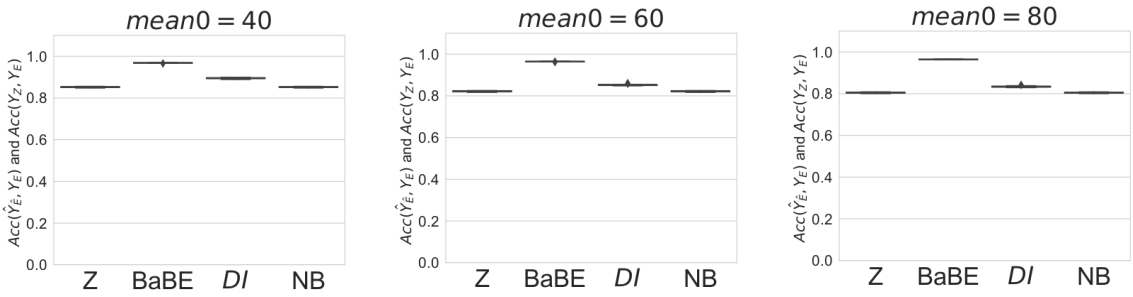


Figure 7.18: The Wasserstein distance between $\hat{\mathbb{P}}[Z]$ and $\mathbb{P}[E]$ and between $\hat{\mathbb{P}}[E]$ and $\mathbb{P}[E]$.

Figures7.19 shows the accuracy for the two groups. Once again the performance of BaBE is better than other pre-processing methods.

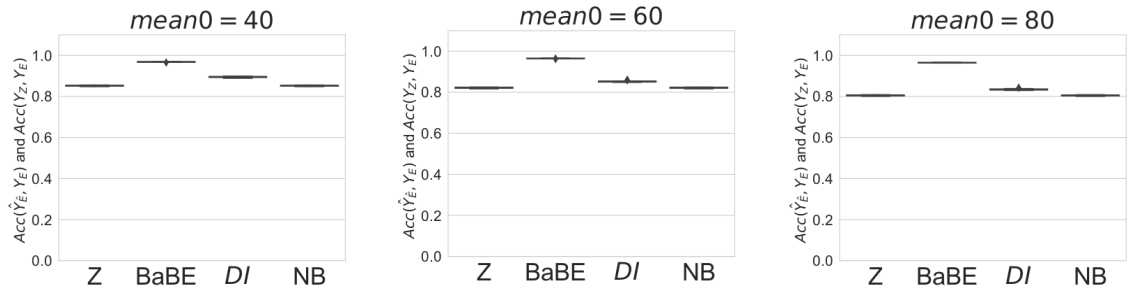


Figure 7.19: Experiment on the transfer of knowledge: The accuracy between $\hat{Y}_Z$ and Y_E (for Z), and between $\hat{Y}_{\hat{E}}$ and Y_E.

Figure 7.20 shows the distortion (Equation 6.3). Babe again produces results that are closer to the true values of E than the ones produced by other methods.

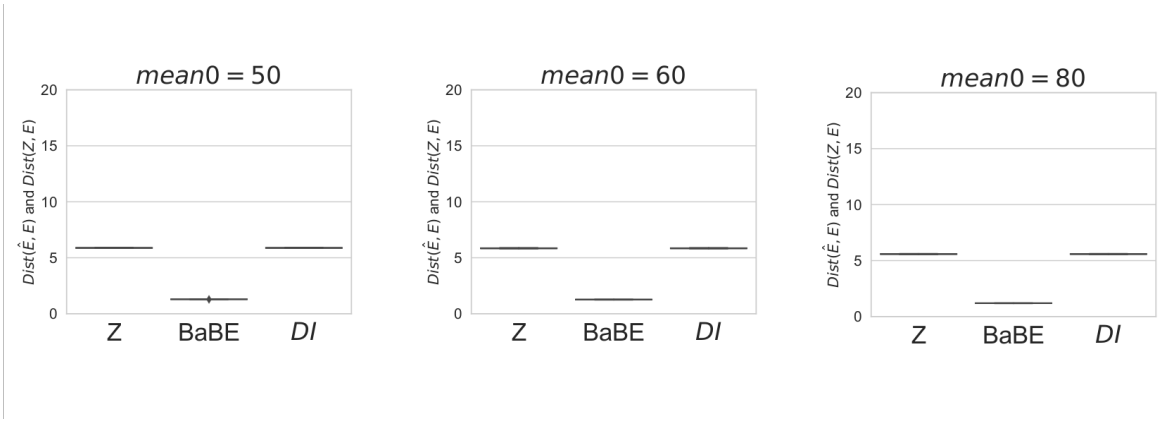


Figure 7.20: The distortion.

Figure 7.21 shows the conditional statistical parity difference on admission for each group, conditioned on E. The values for BaBE are close to zero, indicating the absence of discrimination.

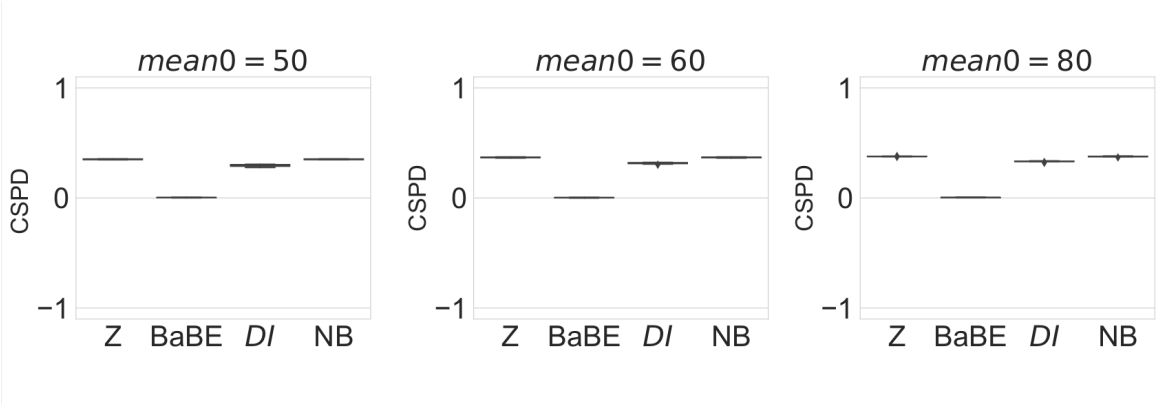


Figure 7.21: Experiment on the transfer of knowledge: Conditional Statistical Parity Difference (CSPD). We recall that for BaBE, DI and NB, CSPD is defined as $\mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|E, S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z.

Figure 7.22 shows the probabilities of positive prediction when the true decision is positive,

and the corresponding difference in equal opportunity. We note that the prediction based on Z has a high probability to be positive for the group 1, but not for the group 0, therefore Z has positive values for EOD. On the other hand, BaBE's prediction is based on the estimation of E, and hence tends to be equal to the true decision yielding EOD close to zero.

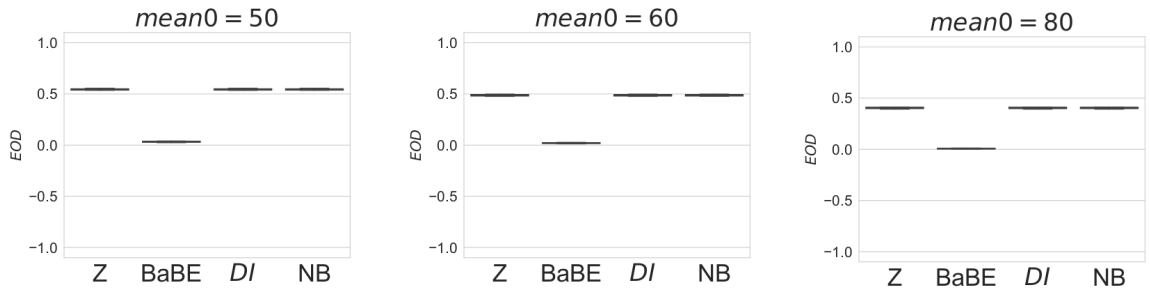


Figure 7.22: Experiment on the transfer of knowledge: Equal Opportunity Difference (EOD). We recall that for BaBE, DI, and NB, EOD is defined as $\mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1] - \mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z

Finally, Figure 7.23 compares the statistical difference (SPD) of the prediction $\hat{Y}_{\hat{E}}$ obtained with the various methods and the SPD of Y_Z. The SPD for $\hat{Y}_{\hat{E}}$ is defined in (7.1), for Y_Z is defined as $\mathbb{P}[Y_Z = 1|S = 1] - \mathbb{P}[Y_Z = 1|S = 0]$. Once again, the SPD of Y_Z decreases as *mean0* (the merit of group 0) increases. BaBE achieves correctly $SPD = 0$ where *mean0* = 60, that is, the same as *mean1*. DI and NB achieve SPD close to zero only in the data with *mean0* = 80, where the disparity is high and the positive decision is more likely for $S = 0$. In the cases where a positive decision is more likely for $S = 1$ (discrimination against $S = 0$), DI and NB tend to overcompensate, resulting in a negative value for SPD.

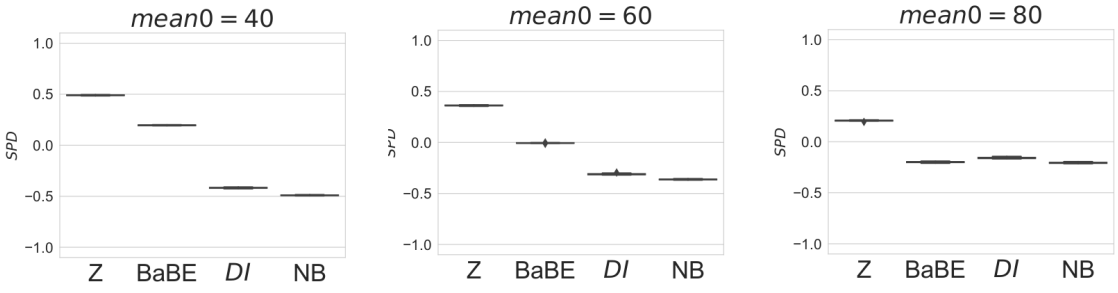


Figure 7.23: Statistical Parity Difference (SPD).

Results on the NHANES data

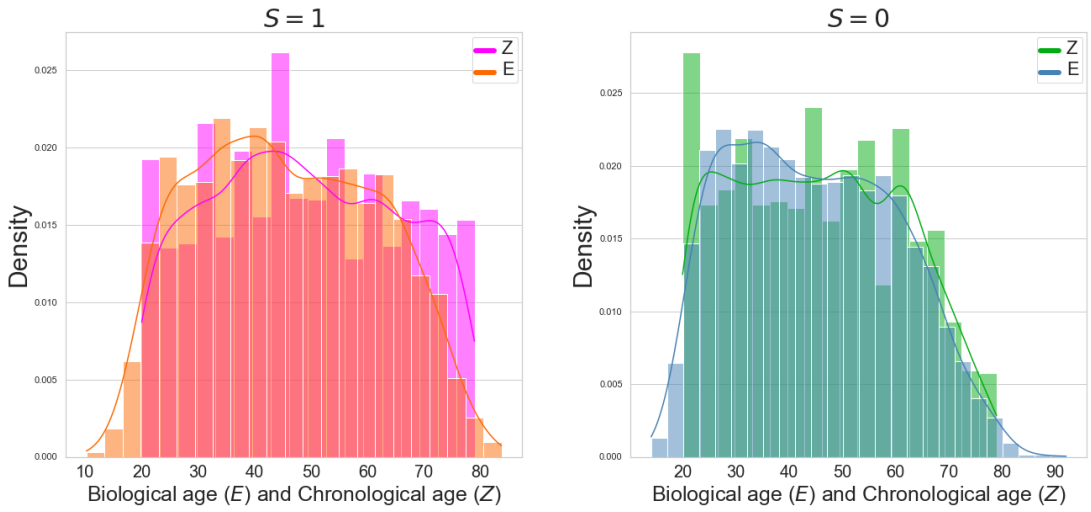


Figure 7.24: Distributions of E and Z for $S = 1$ (left) and $S = 0$ (right) in NHANES data set.

In this section, we show statistics and plots for the NHANES dataset. We applied BaBE using Method 2 (7.3.4), because the conditional distribution of $Z|E, S$ does not allow the accurate estimation of every individual E. However, it still allows us to recover the aggregated distribution and estimate $\hat{Y}_{\hat{E}}$. Consistent with method 2 we report only EOD [8] and $Acc(\hat{Y}_{\hat{E}}, Y_E)$, $Acc(Y_Z, Y_E)$ [9].

Figure 7.25 shows the accuracy resulting from the application of BaBE, DI, and NB to the NHANES data set when the threshold is set to 75 years of age or older. BaBE achieves better overall accuracy and significantly better accuracy for $S = 1$.

Figure 7.26 shows the equal opportunity from the application of BaBE, DI, and NB to the NHANES data set. BaBE achieves EOD close to zero. DI and NB preprocessing methods do not differ significantly from the estimated EOD considering the original Z. DI and NB are designed to aim for a statistical disparity that is equal to zero. The statistical disparity in the NHANES data is very small (owing to the oversampling of the minority population), so much preprocessing is not needed if the goal is optimizing statistical parity.

[8] We also report intermediate steps for EOD: $\mathbb{P}[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1]$ and $\mathbb{P}[Y_Z = 1|Y_E = 1, S = 1]$, $\mathbb{P}[\hat{Y}_{\hat{E}} = |Y_E = 1, S = 0]$ and $\mathbb{P}[Y_Z = 1|Y_E = 1, S = 0]$

[9] In addition we report intermittent steps to obtain accuracy measure: $Acc(\hat{Y}_{\hat{E}}|S = 1, Y_E|S = 1)$ and $Acc(Y_Z|S = 1, Y_E|S = 1)$, $Acc(\hat{Y}_{\hat{E}}|S = 1, Y_E|S = 0)$ and $Acc(Y_Z|S = 1, Y_E|S = 0)$

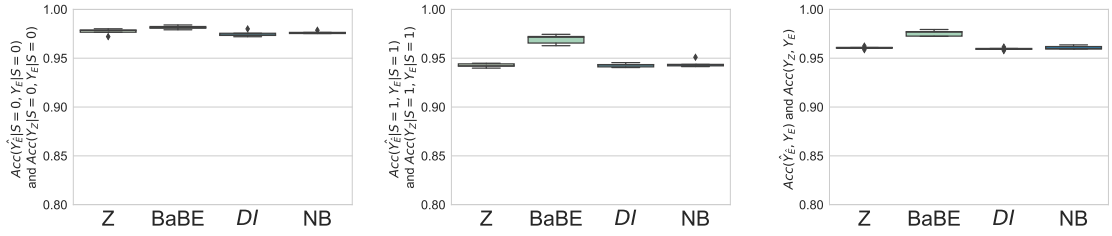


Figure 7.25: Experiments on the NHANES data. The accuracy between $\hat{Y}_Z$ and Y_E (for Z), and between $\hat{Y}_{\hat{E}}$ and Y_E.

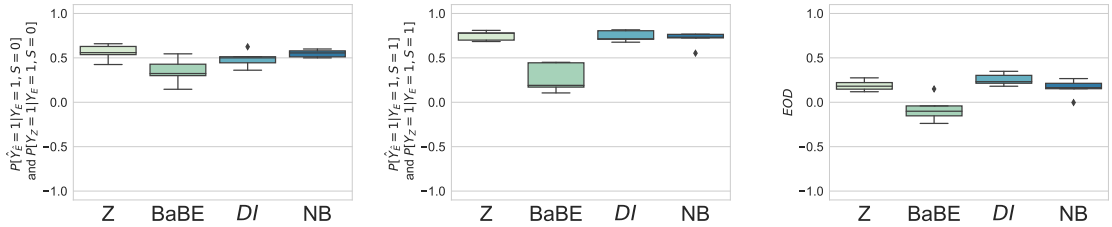


Figure 7.26: Experiments on the NHANES data. EOD.

7.4.1 Discussion

Our experiments show that BaBE performs well for the fairness notions for which BaBE is designed, i.e., CSPD_e and EOD while maintaining good accuracy.

BaBE performs well also when $\mathbb{P}[E|S]$ is different from that of the data in which $\mathbb{P}[Z|E,S]$ has been computed (Figure 7.18), which shows that BaBE is compatible with the transfer of causal knowledge to populations with different distributions. On the contrary, DI and NB highly depend on the distribution as they always aim to minimize SPD. Note that minimizing SPD in the NHANES data set would still result in discrimination against black people, who on average have higher biological age than white people of the same chronological age.

It is important to mention that the performance of BaBE is dependent on the invertibility of $\mathbb{P}[Z|E,S=s]$ (seen as stochastic matrix, aka bias matrix), because invertibility is necessary for the uniqueness of the MLE. However, even when the matrix is not invertible, we are able to obtain favorable results. Indeed, in all our experiments the bias matrices we produce from the synthetic data are not invertible, to mimic the more realistic scenarios. Preliminary experiments show that the diagonal deterministic matrix produces the highest precision for the estimation of distributions $\mathbb{P}[E|S]$, and highest accuracy of the prediction $\hat{Y}_{\hat{E}}$. We leave a more systematic study on how precision and accuracy depend on $\mathbb{P}[Z|E,S=s]$ as a topic for future work.

8

Underrepresentation and Sampling Bias in Machine Learning

8.1 Introduction

A common category of bias in ML occurs when the ML model is trained using a limited number of samples. This produces an inaccurate model and the inaccuracy will typically be born differently by different subpopulations, which leads to discrimination. This category of bias is inconsistently given various names in the literature (e.g. sampling bias, representation bias, data imbalance bias, etc.) and, to the best of our knowledge, is not formally defined. This chapter is an attempt to disambiguate this category of bias by proposing definitions of two sources of bias, namely, sample size bias (SSB) and underrepresentation bias (URB). SSB is the bias that results from training an ML model using training data with a limited number of samples and where all subpopulations are represented in the same proportions as the real population. URB is the bias resulting from training an ML model using training data with a disparity in the number of samples corresponding to each subpopulation.

8.1.1 Related Work

Although the link between the limited number of samples used for training and the disparity in the accuracy of the obtained model may seem straightforward, the magnitude of such a pattern has not been thoroughly studied in the ML fairness literature. Based on the proposed definitions of SSB and URB, the empirical part of the chaper illustrates how the magnitude of discrimination behaves as more extreme versions of bias are considered. Various discrimination metrics are used, namely, difference in FPR (false positive rate), equal opportunity [14], difference in ZOL

(zero-one loss), difference in AUC (area under the curve), statistical disparity [96], and, for regression problems, difference in MSE (mean squared error). For the latter, we use previous results in the literature [153, 154] to decompose discrimination into noise, bias, and variance.

The literature, particularly related to computer vision [6, 155, 156] suggests that sampling bias can be corrected using more data for training, in particular for underrepresented groups. Obtaining more data is possible either through data augmentation (duplicating or creating synthetic samples) or resuming data collection. Unlike data augmentation, whose effect on discrimination has been the subject of a number of papers, in particular related to computer vision (e.g. [157–163]), the impact of collecting more samples on discrimination has not been well studied in the literature[1]. Furthermore, the effect of collecting more samples on discrimination in the case low-dimentional tabular data has not been addressed.

8.2 Preliminaries

Let $\mathcal{A}$ be a supervised learning algorithm for learning an unknown function $f : \boldsymbol{\mathcal{X}} \mapsto \mathcal{Y}$ where $\boldsymbol{\mathcal{X}}$ is the input variables space and $\mathcal{Y}$ is the outcome space. Without loss of generality, the outcome random variable Y is assumed to be binary ($\mathcal{Y} = \{0, 1\}$, e.g. accepted/rejected). Let $\mathcal{S} = \{(\mathbf{x}_i, y_i = f(\mathbf{x}_i))\}, i = 1 \dots m$, be a training sample of size m. Based on the data sample $\mathcal{S}$, algorithm $\mathcal{A}$ learns a function $\mathcal{A}(\mathcal{S}) = \hat{f}_{\mathcal{S}}^{\mathcal{A}}$. Let $\hat{Y}_{\mathcal{S}}^{\mathcal{A}}$ be the predicted outcome random variable such that $\hat{f}_{\mathcal{S}}^{\mathcal{A}}(\mathbf{x}_i) = \hat{y}_i$. When there is no ambiguity, we refer to $\hat{Y}_{\mathcal{S}}^{\mathcal{A}}$ and $\hat{f}_{\mathcal{S}}^{\mathcal{A}}$ simply as $\hat{Y}$ (or $\hat{Y}_{\mathcal{S}}$) and $\hat{f}$ (or $\hat{f}_{\mathcal{S}}$).

Given the true value y and the prediction $\hat{y}$, $L(y, \hat{y})$ represents the loss incurred by predicting $\hat{y}$ while the true outcome is y. A commonly used loss function for regression problems is the squared loss defined as $L^{SL}(\hat{y}, y) = (\hat{y} - y)^2$. Other loss functions that will be considered in this chapter are the absolute loss $L^{AL}(\hat{y}, y) = |\hat{y} - y|$ and the zero-one loss $L^{ZO}(\hat{y}, y) = 0$ if $\hat{y} = y$, and 1 otherwise.

Based on a loss function, we define two special predictions, namely, the main prediction for a learning algorithm $\mathcal{A}$ and the optimal prediction.

Given a learning algorithm $\mathcal{A}$ and a set of training samples $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots\}$, the main prediction random variable $\bar{\hat{Y}}_{\mathfrak{S}}^{\mathcal{A}}$ ($\bar{\hat{y}} = \bar{\hat{f}}_{\mathfrak{S}}^{\mathcal{A}}(\mathbf{x})$) represents the prediction that minimizes the loss across all training sets in $\mathfrak{S}$. That is,

$$\bar{\hat{f}}_{\mathfrak{S}}^{\mathcal{A}}(\mathbf{x}) = \underset{f'}{\operatorname{argmin}}\, \mathbb{E}_{\mathcal{S}\in\mathfrak{S}}[L(\hat{f}_{\mathcal{S}}(\mathbf{x}), f'(\mathbf{x})].$$

When there is no ambiguity, we refer to $\bar{\hat{Y}}_{\mathfrak{S}}^{\mathcal{A}}$ and $\bar{\hat{f}}_{\mathfrak{S}}^{\mathcal{A}}(\mathbf{x})$ simply as $\bar{\hat{Y}}$ and $\bar{\hat{f}}(\mathbf{x})$. Typically, the main prediction corresponds to the average prediction across all training sets in $\mathfrak{S}$. That is,

$$\bar{\hat{f}}(\mathbf{x}) = \underset{\mathcal{S}\in\mathfrak{S}}{\mathbb{E}} \hat{f}_{\mathcal{S}}(\mathbf{x})^2. \tag{8.1}$$

The optimal prediction Y^* ($y^* = f^*(\mathbf{x})$) is the prediction that minimizes the loss across all

[1] Detailed related work is provided in Supplementary Material

[2] We exceptionally use the expectation on a function, instead of a random variable.

possible predictors. That is,

$$f^*(\mathbf{x}) = \underset{f'}{\operatorname{argmin}}\, \mathbb{E}[L(f(\mathbf{x}), f'(\mathbf{x})].$$

It is important to note that f^* is independent of the learning algorithm $\mathcal{A}$.

Assume that the sensitive attribute A is a binary variable with possible values $A = a_0$ and $A = a_1$, each representing a different group (e.g. male vs female, black vs white, etc.). Let G_0 and G_1 denote these groups. That is, $G_0 = \{\mathbf{x} \in \mathcal{X} | A = a_0\}$ and $G_1 = \{\mathbf{x} \in \mathcal{X} | A = a_1\}$. Discrimination between G_0 and G_1 can be defined in terms of the disparity in prediction accuracy. Let $C_a^\bullet(\hat{Y})$ denote the accuracy/cost of prediction $\hat{Y}$ for group $A = a$. For classification problems, we consider four metrics, namely, false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), and zero one loss (ZOL). For regression problems, we consider mean square error (MSE). These metrics are defined as follows:

- $C_a^{FPR}(\hat{Y}) = \mathbb{E}[\hat{Y} | Y = 0, A = a]$
- $C_a^{FNR}(\hat{Y}) = \mathbb{E}[1 - \hat{Y} | Y = 1, A = a]$
- $C_a^{TPR}(\hat{Y}) = \mathbb{E}[\hat{Y} | Y = 1, A = a]$
- $C_a^{ZOL}(\hat{Y}) = \mathbb{E}[\mathbb{1}[\hat{Y} \neq Y] | A = a]$
- $C_a^{MSE}(\hat{Y}) = \mathbb{E}[(\hat{Y} - Y)^2 | A = a]$

Discrimination $Disc^\bullet$ can be defined as the difference in $C_a^\bullet$ between the two sensitive groups. For instance $Disc^{FPR}(\hat{Y}) = C_{a_1}^{FPR}(\hat{Y}) - C_{a_0}^{FPR}(\hat{Y})$. Notice that $Disc^{TPR}(\hat{Y})$ corresponds to discrimination according to equal opportunity [14] and that $Disc^{TPR}(\hat{Y}) = -Disc^{FNR}(\hat{Y})$ as $TPR = 1 - FNR$. In the rest of the chapter, we use $Disc^{TPR}(\hat{Y})$ and $Disc^{EO}(\hat{Y})$ interchangeably. In addition, for reference, we use $Disc^{SD}(\hat{Y}) = \mathbb{E}[\hat{Y} | A = a_1] - \mathbb{E}[\hat{Y} | A = a_0]$ to denote statistical disparity [96].

8.2.1 Decomposing and bounding statistical disparity

Statistical disparity is the simplest discrimination metric and it corresponds to the difference in the expected outcomes between groups:

Definition 8.2.1 (Statistical Disparity)**.**

$$\begin{aligned} Disc^{SD}(\hat{Y}_S) &= \mathbb{E}_X[\hat{Y}_S | A = a_1] - \mathbb{E}_X[\hat{Y}_S | A = a_0] \\ &= \mathop{\mathbb{E}}_{\mathbf{x} \in G_1}[\hat{f}_S(\mathbf{x})] - \mathop{\mathbb{E}}_{\mathbf{x} \in G_0}[\hat{f}_S(\mathbf{x})] \end{aligned}$$

$Disc^{SD}(\hat{Y}_S)$ is a biased estimation of the *true* value $Disc^{SD}(Y)$. The following theorem states that the error in estimating statistical disparity can be bounded where the bounds are expressed in terms of noise, bias, and variance.

THEOREM 8.1. The error in estimating statistical disparity is bounded as follows:

$$|Disc^{SD}(\hat{Y}_S) - Disc^{SD}(Y)| \leq (\overline{N}^{AL}_{a_1}(\hat{Y}_S) - \overline{N}^{AL}_{a_0}(\hat{Y}_S)) + (\overline{B}^{AL}_{a_1}(\hat{Y}_S) - \overline{B}^{AL}_{a_0}(\hat{Y}_S)) + (\overline{V}^{AL}_{a_1}(\hat{Y}_S) - \overline{V}^{AL}_{a_0}(\hat{Y}_S))$$

$$\begin{aligned}|Disc^{SD}(\hat{Y}_S) - Disc^{SD}(Y)| \geq \max(& \\ & (\overline{N}^{AL}_{a_1}(\hat{Y}_S) - \overline{N}^{AL}_{a_0}(\hat{Y}_S)) - (\overline{B}^{AL}_{a_1}(\hat{Y}_S) - \overline{B}^{AL}_{a_0}(\hat{Y}_S)) - \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_S) - \overline{V}^{AL}_{a_0}(\hat{Y}_S)), \\ & (\overline{B}^{AL}_{a_1}(\hat{Y}_S) - \overline{B}^{AL}_{a_0}(\hat{Y}_S)) - (\overline{N}^{AL}_{a_1}(\hat{Y}_S) - \overline{N}^{AL}_{a_0}(\hat{Y}_S)) - \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_S) - \overline{V}^{AL}_{a_0}(\hat{Y}_S)), \\ & (\overline{V}^{AL}_{a_1}(\hat{Y}_S) - \overline{V}^{AL}_{a_0}(\hat{Y}_S)) - (\overline{B}^{AL}_{a_1}(\hat{Y}_S) - \overline{B}^{AL}_{a_0}(\hat{Y}_S)) - \\ & (\overline{N}^{AL}_{a_1}(\hat{Y}_S) - \overline{N}^{AL}_{a_0}(\hat{Y}_S)))\end{aligned}$$

where:

- $\overline{N}^{AL}_{a}(\hat{Y}_S) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[N^{AL}(\mathbf{x})|A = a]$
- $\overline{B}^{AL}_{a}(\hat{Y}_S) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[B^{AL}(\mathbf{x})|A = a]$
- $\overline{V}^{AL}_{a}(\hat{Y}_S) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[(1 - 2 \times B^{AL}(\mathbf{x})) \times V^{AL}(\mathbf{x})|A = a]$

Proof. The proof is based on the triangle inequality of metrics. Recall that a metric is a function of two arguments ($dist(x, y)$) that satisfy minimality ($\forall x, y, dist(x, y) \geq dist(x, y)$), symmetry ($\forall x, y, dist(x, y) = dist(y, x)$), and triangle inequality ($\forall x, y, z dist(x, z) + dist(z, x) \geq dist(x, y)$). The full proof is very similar to the proof in [154] (Theorem 7). □

8.3 Sample Size and Underrepresentation Biases

Typically, the size of the data used to train an ML model has a significant impact on the accuracy of the obtained model. However, it is generally assumed that the loss in accuracy is equally born by the different segments of the data. As it is not usually the case, we define sample size bias (SSB) as the bias resulting from training a model with a given data size.

Let $\mathfrak{S}_m = \{\mathcal{S}_1, \mathcal{S}_2, \ldots\}$ be the set of samples of size m, and let $\hat{f}_{\mathcal{S}_1}, \hat{f}_{\mathcal{S}_2}, \ldots$ be the models produced by applying the learning algorithm $\mathcal{A}$ on each sample ($\mathcal{A}(\mathcal{S}_1) = \hat{f}_{\mathcal{S}_1}$, etc.). Let $\bar{\hat{Y}}^{\mathcal{A}}_{\mathfrak{S}_m}$ ($\bar{\hat{y}}_m = \bar{\hat{f}}^{\mathcal{A}}_{\mathfrak{S}_m}(\mathbf{x})$) be the main prediction obtained using the set of training sets $\mathfrak{S}_m$. That is,

$$\bar{\hat{f}}^{\mathcal{A}}_{\mathfrak{S}_m}(\mathbf{x}) = \underset{f'}{\operatorname{argmin}}\, \mathbb{E}_{\mathcal{S}\in\mathfrak{S}_m}[L(\hat{f}_{\mathcal{S}}(\mathbf{x}), f'(\mathbf{x})]. \tag{8.2}$$

When there is no ambiguity, we refer to $\bar{\hat{Y}}^{\mathcal{A}}_{\mathfrak{S}_m}$ and $\bar{\hat{f}}^{\mathcal{A}}_{\mathfrak{S}_m}$ simply as $\bar{\hat{Y}}$ and $\bar{\hat{f}}_m$.

DEFINITION 8.3.1. Given a positive number $m > 0$ representing the training set size, sample size bias is the difference in discrimination due to the training set

size:

$$SSB^{\bullet}(\mathcal{A}, m) = Disc^{\bullet}(\bar{\hat{Y}}_m) - Disc^{\bullet}(\bar{\hat{Y}}_\infty) \tag{8.3}$$

where $Disc^{\bullet}(\bar{\hat{Y}}_\infty) = \lim_{m\to\infty} Disc^{\bullet}(\bar{\hat{Y}}_m)$ and $\bullet$ is a placeholder for the accuracy/cost metric (FPR, FNR, EO, ZOL, or MSE for regression problems). As a metric that combines both specificity (FPR) and sensitivity (TPR), we use also AUC (area under the curve)[3]. For reference, we consider also statistical disparity that we denote as $Disc^{SD}$ (See Supplementary Material).

As SSB is defined in terms of an infinite size training set ($\bar{\hat{Y}}_\infty$), we consider an alternative definition in terms of M, the size of the largest training set available:

$$SSB^{\bullet}_{M}(\mathcal{A}, m) = Disc^{\bullet}(\bar{\hat{Y}}_m) - Disc^{\bullet}(\bar{\hat{Y}}_M) \tag{8.4}$$

Another variant of SSB can be defined based on a specific training set $\mathcal{S}_m$ of size m as follows:

$$SSB_M{}^{\bullet}(\mathcal{A}, \mathcal{S}_m) = Disc^{\bullet}(\hat{Y}_{\mathcal{S}_m}) - Disc^{\bullet}(\bar{\hat{Y}}_M) \tag{8.5}$$

When sampling a training set from a population, it is generally assumed that the generated sample is balanced. Data are balanced if all classes are proportionally represented and are imbalanced if they suffer from severe class distribution skews [164]. For instance, if a class label is overrepresented at the expense of another underrepresented class label. If data is imbalanced in the sensitive groups (e.g. male vs female, blacks vs whites, etc.), it can have significant impact on the disparity of accuracies and consequently on discrimination between sensitive groups. We define underrepresentation bias (URB) as the bias resulting from a disparity in representation between the sensitive groups.

Let $\mathfrak{S}_m^{\frac{m_1}{m_0}}$ be the set of samples of size m with m_0 and m_1 items from G_0 and G_1 respectively. That is, for $\mathcal{S} \in \mathfrak{S}_m^{\frac{m_1}{m_0}}$, $|\{\mathbf{x} \in \mathcal{S} | A = a_0\}| = m_0$, $|\{\mathbf{x} \in \mathcal{S} | A = a_1\}| = m_1$, and $m_0 + m_1 = m = |\mathcal{S}|$. We use the simpler notation $\bar{\hat{Y}}_{\frac{m_1}{m_0}}$ to refer to $\bar{\hat{Y}}^{\mathcal{A}}_{\mathfrak{S}_m^{\frac{m_1}{m_0}}}$.

Definition 8.3.2. Given, $m, m_0, m_1 > 0$ such that $m_0 + m_1 = m$, underrepresentation bias is the difference in discrimination due to the disparity in sample sizes compared to the population ratio:

$$URB^{\bullet}(\mathcal{A}, m_0, m_1) = Disc^{\bullet}(\bar{\hat{Y}}_{\frac{m_1}{m_0}}) - Disc^{\bullet}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) \tag{8.6}$$

where $Disc^{\bullet}(\bar{\hat{Y}}_{m_1^p/m_0^p})$ is the discrimination of the prediction based on a model trained using only samples from $\mathfrak{S}_m^{m_1^p/m_0^p}$, and the ratio $\frac{m_1^p}{m_0^p}$ is the same as the ratio in the population ($\frac{m_1^p}{m_0^p} \approx \frac{|G_1|}{|G_0|}$).

Similar to $SSB^{\bullet}_{M}(\mathcal{A}, \mathcal{S}_m)$ (Equation 8.5), a variant of URB can be defined based on a specific

[3]Other metrics combining specificity and sensitivity include F_1 score and balanced accuracy (BA)

training set $\mathcal{S}_{\frac{m_1}{m_0}} \in \mathfrak{S}_m^{\frac{m_1}{m_0}}$ as follows:

$$URB^{\bullet}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}}) = Disc^{\bullet}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - Disc^{\bullet}(\bar{\hat{Y}}_{m_1^P \atop m_0^P}) \tag{8.7}$$

8.4 Loss and Discrimination Decomposition

Domingos [154] showed that if a learning algorithm $\mathcal{A}$ learns a function $\mathcal{A}(\mathcal{S}) = \hat{f}_{\mathcal{S}}$ based on a training set $\mathcal{S} \in \mathfrak{S}$, then the expected loss between the prediction $\hat{f}_{\mathcal{S}}(\mathbf{x})$ and the true value $f(\mathbf{x})$ can be decomposed into noise, bias, and variance. In particular, for squared loss,

$$L^{SL}(\hat{f}_{\mathcal{S}}(\mathbf{x}), f(\mathbf{x})) = N^{SL}(\mathbf{x}) + B^{SL}(\mathbf{x}) + V^{SL}(\mathbf{x}) \tag{8.8}$$

where

- $N^{SL}(\mathbf{x}) = L^{SL}(f^*(\mathbf{x}), f(\mathbf{x}))$
- $B^{SL}(\mathbf{x}) = L^{SL}(\bar{\hat{f}}(\mathbf{x}), f^*(\mathbf{x}))$
- $V^{SL}(\mathbf{x}) = \mathbb{E}_{\mathcal{S}\in\mathfrak{S}}[L^{SL}(\hat{f}_{\mathcal{S}}(\mathbf{x}), \bar{\hat{f}}(\mathbf{x}))]$

The loss decomposition can be illustrated as follows:

$$f(\mathbf{x}) \underset{Noise}{\longleftrightarrow} f^*(\mathbf{x}) \underset{Bias}{\longleftrightarrow} \bar{\hat{f}}(\mathbf{x}) \underset{Variance}{\longleftrightarrow} \hat{f}_{\mathcal{S}}(\mathbf{x})$$

For Zero-One loss (L^{ZO}), Equation 8.8 holds also but with coefficients different than 1 for the noise and variance terms. However, it does not hold for the absolute loss (L^{AL})[4] [154].

8.4.1 Decomposing Discrimination

Chen et al. [153] showed that the accuracy/cost metric $C_a^{\bullet}(\hat{Y}_{\mathcal{S}})$ as well as the discrimination $Disc^{\bullet}(\hat{Y}_{\mathcal{S}})$ can be decomposed into noise, bias, and variance components. In particular, for MSE,

$$C_a^{MSE}(\hat{Y}_{\mathcal{S}}) = \overline{N}_a^{SL}(\hat{Y}_{\mathcal{S}}) + \overline{B}_a^{SL}(\hat{Y}_{\mathcal{S}}) + \overline{V}_a^{SL}(\hat{Y}_{\mathcal{S}}) \tag{8.9}$$

where:

- $\overline{N}_a^{SL}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[N^{SL}(\mathbf{x})|A = a]$
- $\overline{B}_a^{SL}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[B^{SL}(\mathbf{x})|A = a]$
- $\overline{V}_a^{SL}(\hat{Y}_{\mathcal{S}}) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}[(1 - 2 \times B^{SL}(\mathbf{x})) \times V^{SL}(\mathbf{x})|A = a]$

The last term ($V_a(\hat{Y}_{\mathcal{S}})$) is called *net variance* [154]. Consequently,

$$Disc^{MSE}(\hat{Y}_{\mathcal{S}}) = (\overline{N}_{a_1}^{SL}(\hat{Y}_{\mathcal{S}}) - \overline{N}_{a_0}^{SL}(\hat{Y}_{\mathcal{S}})) + (\overline{B}_{a_1}^{SL}(\hat{Y}_{\mathcal{S}}) - \overline{B}_{a_0}^{SL}(\hat{Y}_{\mathcal{S}})) + (\overline{V}_{a_1}^{SL}(\hat{Y}_{\mathcal{S}}) - \overline{V}_{a_0}^{SL}(\hat{Y}_{\mathcal{S}})) \tag{8.10}$$

[4] Alternatively, upper and lower bounds are possible.

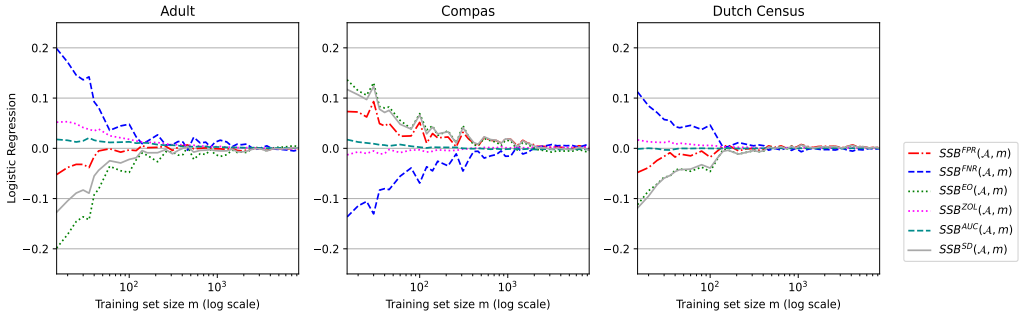


Figure 8.1: Magnitude of sample size bias (SSB) for increasing size of the training data.

The decomposition of Equation 8.10 will also hold for $Disc^{FPR}(\hat{Y}_S)$, $Disc^{EO}(\hat{Y}_S)$, and $Disc^{ZOL}(\hat{Y}_S)$ but with coefficients different than 1 for the noise and variance terms [153].

8.4.2 Decomposing SSB and URB

The variant $SSB^{\bullet}_{M}(\mathcal{A}, \mathcal{S}_m)$ (Eq. 8.5) of sample size bias has the advantage that it can be decomposed into bias and variance. The decomposition for the MSE metric is as follows.

THEOREM 8.2. *$SSB^{MSE}_{M}(\mathcal{A}, \mathcal{S}_m)$ can be decomposed into bias and variance components as follows:*

$$\begin{aligned} SSB_M{}^{MSE}(\mathcal{A}, \mathcal{S}_m) = \overline{B}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_m}) - \overline{B}^{SL}_{a_1}(\bar{\hat{Y}}_M) - (\overline{B}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_m}) - \overline{B}^{SL}_{a_0}(\bar{\hat{Y}}_M)) \\ + \overline{V}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_m}) - \overline{V}^{SL}_{a_1}(\bar{\hat{Y}}_M) - (\overline{V}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_m}) - \overline{V}^{SL}_{a_0}(\bar{\hat{Y}}_M)) \end{aligned} \tag{8.11}$$

Proof. The proof follows from Equation 8.9 and from assuming that the optimal predictor Y^* coincides with the true value Y and hence noise is 0 [5]. □

$URB^{\bullet}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}})$ (Equation 8.7) can also be decomposed into bias and variance components. The decomposition for the MSE metric is as follows.

THEOREM 8.3.

$$\begin{aligned} URB^{MSE}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}}) = \overline{B}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{B}^{SL}_{a_1}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) - (\overline{B}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{B}^{SL}_{a_0}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}})) \\ + \overline{V}^{SL}_{a_1}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{V}^{SL}_{a_1}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}}) - (\overline{V}^{SL}_{a_0}(\hat{Y}_{\mathcal{S}_{\frac{m_1}{m_0}}}) - \overline{V}^{SL}_{a_0}(\bar{\hat{Y}}_{\frac{m_1^p}{m_0^p}})) \end{aligned}$$

Proof. The same as Theorem 8.2. □

8.5 Experimental Analysis

The objective of the experimental analysis is to observe the magnitude of both types of bias, namely sample size bias SSB and underrepresentation bias URB as we change the parameters

[5] We follow previous work (Domingos [154] and Kohavi and Wolpert [165]) in assuming a zero noise.

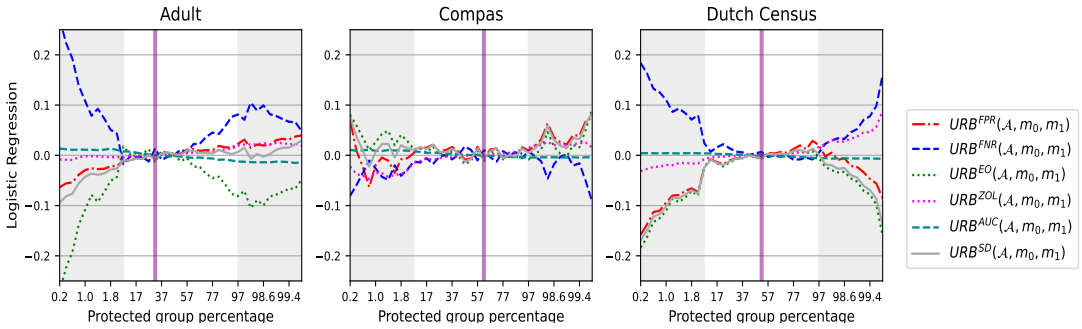


Figure 8.2: Underrepresentation Bias (URB) for different ratios of sensitive groups. The size of the training set is fixed (1000). The horizontal bar represents the same ratio as the population. The shaded sections indicate a focus on the extreme proportions (less than 2% and more than 98%)

of data sampling. For SSB, we train the predictor model using training sets of increasing size. For URB, we play rather on the proportions of sensitive groups in the training set. In this case, we have set the sample size to a minimum number of instances shown to be stable in terms of sampling bias in the previous experiments. In this way, we decouple the underrepresentation bias from the sampling bias. Three benchmark datasets are used, Adult [166], Compas [167], and Dutch Census [168][6].

8.5.1 Magnitude of sample size bias (SSB)

To observe how sample size bias behaves as the training set size changes, we use the following process. We use a sequence of sample sizes ranging from 10 until a given portion of the full size of the data set. For example, for COMPAS, we consider sample sizes ranging from 10 to 2000. For each sample size value m, we repeat the sampling several times (30 by default) so that we obtain 30 samples of each size m. Then, we train a different model using each one of the samples so that we obtain 30 models for each size m. We finally compute the discrimination using each model, and the returned value is the average discrimination across all models. This procedure gives a sequence of discrimination values indexed by the size. We consider five cost / accuracy metrics, namely FPR (false positive rate), FNR (false negative rate), EO (equal opportunity), ZOL (zero one loss), and SD (statistical disparity). We use five classifiers, namely logistic regression, decision tree, random forest, nearest neighbor, and support vector machine (SVM).

Figure 8.1 shows the magnitude of SSB according to each metric and for each benchmark data set and using logistic regression. Notice that SSB^{EO} and SSB^{FNR} are symmetric because, as mentioned above, $FNR = 1 - TPR$ and hence $SSB^{EO} = -SSB^{FNR}$. Most of the plots exhibit an expected behavior of SSB. That is, the bias is significant when the models are trained using a limited-size training set. The bias disappears gradually as the size of the training set increases. SSB behaves the same way for the other classifiers (Figure 8 in Supplementary Material). More importantly, the results of SSB show that the cost / accuracy metrics that combine specificity and sensitivity (AUC and ZOL) are less sensitive to the size of the training set than the remaining

[6]We use the same dataset versions and learning algorithms parameters as IBM AIF360 [169]

metrics (FPR and EO). A possible explanation is that for small training sets, it is more likely that a majority of the samples have the same outcome (positive or negative), which can boost precision on the expense of recall or the opposite. AUC and ZOL are not subject to this skewness, since they consider the trade-off between precision and recall.

8.5.2 Magnitude of underrepresentation bias (URB)

The aim of the underrepresentation bias experiment is to observe the magnitude of URB while the ratio of the sensitive groups in the training set is changing. We consider different values of the splitting $\frac{m_1}{m_0}$ (see Definition 8.3.2) (e.g. 0.1 vs 0.9, 0.2 vs 0.8, etc.). However, as URB is more significant for extreme disparities, we focus more on extreme splitting values (e.g. 0.001 vs 0.99, 0.002 vs 0.98, etc.). A similar behavior has previously been observed by Farrand et al. [170]. Assuming a fixed sample size (e.g. 1000), for each splitting value, we sample the data so that the proportions of sensitive groups (e.g., male vs. female) match the splitting value. Similarly to the SSB experiment, we repeat the sampling several times (30 by default) for the same splitting value. Next, we train a different model using each of the samples so that we obtain 30 models for each splitting value $\frac{m_1}{m_0}$. The discriminations obtained using the different models are then averaged across all models. We finally obtain a sequence of discrimination values indexed by the splitting value. Figure 8.2 shows how the URB changes as the proportion of the sensitive group increases for the same three data sets and for the use of logistic regression as learning algorithm. The purple vertical bar indicates the percentage of the sensitive group in the entire data set (population). For example, for the adult data set, the percentage of females is 31%. The shaded parts in the background of Figure 8.2's plots indicate that we are "zooming" on the extreme values (the plots are using different steps for the shaded and unshaded parts[7]). Almost all plots exhibit the same pattern for URB, that is, the further the proportions of sensitive groups are from the population proportion reference (vertical bar), the higher the bias. The same expected behavior for URB is obtained when using the other classifiers (Figure 9 in Supplementary Material). The resilience of AUC and ZOL metrics to extreme training set sizes holds also for imbalanced training sets. Notice that URB^{AUC} and URB^{ZOL} remain stable even for extremely imbalanced training sets.

8.5.3 Bias Decomposition

Section 8.4 shows that loss and discrimination can be decomposed into variance, bias, and noise. In particular, assuming that the optimal prediction (Y^*) coincides with the correct outcome (Y), Theorems 8.2 and 8.3 illustrate how $SSB^{MSE}_M(\mathcal{A}, \mathcal{S}_m)$ and $URB^{MSE}(\mathcal{A}, \mathcal{S}_{\frac{m_1}{m_0}})$ can be decomposed into variance and bias components. To illustrate the decomposition empirically, we use the Law School benchmark dataset [3] which tracked some 27 thousand law students through law school and graduation and where the sensitive attribute is gender and the outcome is the first year GPA. We use the scikit-learn linear regression algorithm to train different models using different size training sets. For SSB, the training size m ranges from 10 to $10,000$. For URB, the size

[7]The step is very small below 2% and above 98%.

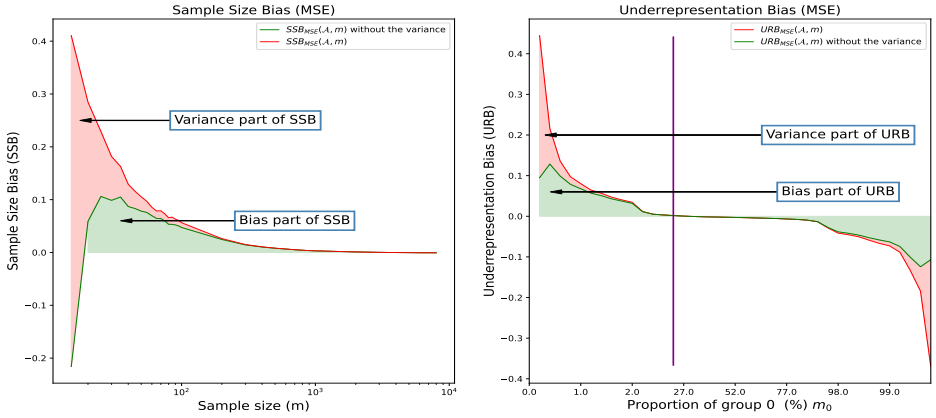


Figure 8.3: Decomposing SSB^{MSE} (left plot) and URB^{MSE} (right plot). The models are trained using linear regression. The benchmark dataset is Law School [3].

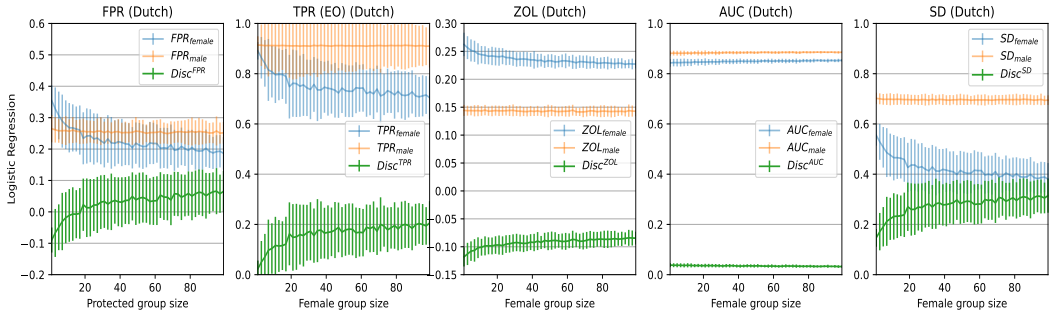


Figure 8.4: Discrimination while augmenting the training set with female group samples randomly. The male group size is fixed at 100. Data set is Dutch Census and training algorithm is logistic regression.

of the training set (m) is fixed at 1000, but the proportion of the protected group (female) is ranging from 0.1% to 99.9%. For each size of training set, the training and testing is repeated 30 times. Figure 8.3 shows how SSB and URB are decomposed into variance and bias. For SSB, the variance component is so significant when the training set is extremely small (less than 20) that it reverses the direction of bias (in favor of females instead of against female). For URB, the variance is also significant when one of the groups is extremely underrepresented, but not to the point of reversing the direction of the bias. The main conclusion from this empirical result is that for very small or very imbalanced training sets, SSB and URB variance can be so important that it can lead to unreliable conclusions about discrimination.

8.5.4 Effect of collecting more samples on discrimination

The natural approach to address sampling bias is to use more data for training, in particular for the under-represented groups. Obtaining more data is possible either through data augmentation or data collection. Data augmentation is the process of using the available data to generate more samples. In turn, this can be done in two ways: oversampling or creating fake samples. Oversampling consists in duplicating existing samples to balance the data. A simple variant is to randomly duplicate samples from the under-represented group. On the other hand, creating

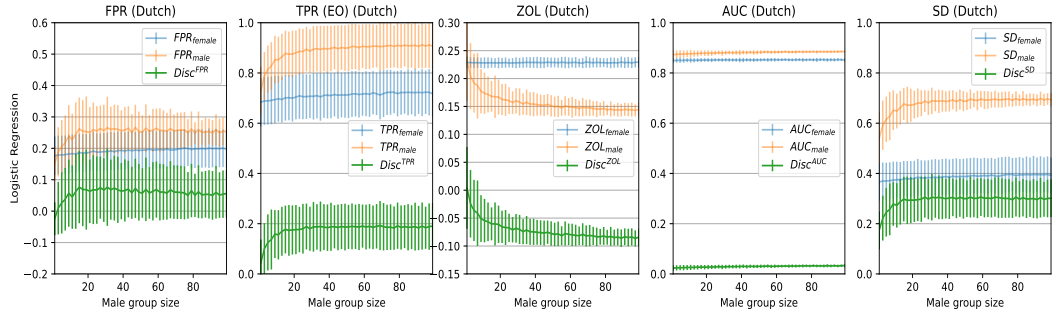


Figure 8.5: Discrimination while augmenting the training set with male group samples randomly. The female group size is fixed at 100. Dataset is Dutch Census and training algorithm is logistic regression.

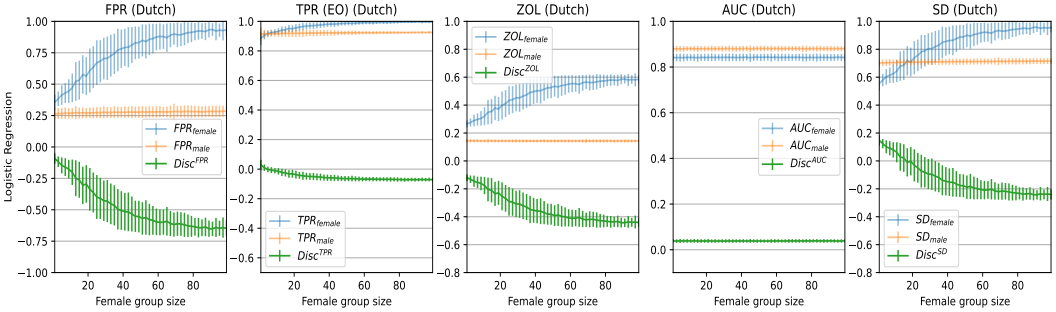


Figure 8.6: Discrimination while augmenting the training set with only positive outcome female group samples. The male group size is fixed at 100. Dataset is Dutch Census and training algorithm is logistic regression.

fake samples is typically done using SMOTE [99]. SMOTE creates synthetic samples based on the k-nearest neighbors of every sample in the under-represented group. Both data-augmentation techniques try to balance data by adding artificially generated samples. Although this artificial manipulation may reduce discrimination between sensitive groups, it can lead to models which are not faithful to reality. When possible, collecting more data is more natural and reflects a better reality. The approach is simple: if a sensitive group is under-represented, collect more samples of that group.

In the following, we devise simple experiments to observe the effect of populating the data with more samples collected from the same population as the existing data. Using the same benchmark datasets, the aim is to train models based on an increasing number of under-represented group samples while keeping the privileged group portion unchanged. For the particular case of Dutch Census dataset, we train models using a set composed of a fixed 100 privileged group (male) samples and an increasing number of protected group (female) samples starting from 2 until 100 (perfect balance between groups). Similarly to the SSB and URB experiments, it turns out that the magnitude of discrimination manifests itself more with extreme values of protected group sizes (typically less than 100) which explains the specific sample sizes considered. Figure 8.4 shows how the cost/accuracy metric values for each group, as well as the corresponding difference (discrimination) change as more protected group samples are considered for model training. We use three-fold cross-validation and since we randomly

generate 50 different samples for every size value, the plots are shown with error bars. As expected, the cost/accuracy metric value for the male group maintains the same mean, while for the female group it is changing. Interestingly, according to all cost/accuracy metrics (except AUC), discrimination is increasing as data is more balanced. Figure 8.4 shows the results with logistic regression, but the pattern is similar for other classification algorithms (Figure 10 in Supplementary Material) and for other benchmark datasets (Appendix B). This counterintuitive behavior is also observed for the reverse experiment where the protected group (female) sample size is fixed (100 samples) while the privileged group (male) is under-represented and more samples are collected and considered in the training (Figure 8.5). It is important to mention that in all previous experiments, selecting samples to balance the training set is performed randomly to simulate, as accurately as possible, data collection in real scenarios. The fairness enhancing potential of adding more samples for the sensitive group depends on the initial fairness characteristics of the data and the goal of the classifier. Wang et al. [161] point out that adding more samples of the minority group to the data increases predictive accuracy and fairness specifically in the classification tasks, where a sensitive attribute is part of the classification output, for example, face recognition [6].

However, if the training set is balanced by selecting a specific type of sample, in particular, samples from the protected group with positive outcome, discrimination will decrease as the data are balanced (Figure 8.6). In all three experiments (collecting more protected group samples randomly, collecting more unprotected group samples randomly, and collecting only protected group samples with positive results), the importance of the sensitive feature (Sex) in the prediction (*shap* explanation [171]) behaves the same way (Figure B.5 in the Appendix B) that is, it contributes more to the learned model as the data is more balanced.

Part IV

Privacy

9

Causal Discovery under Local Privacy

9.1 Introduction

A recent advance in causal discovery is the design of algorithms that estimate the causal structure from observational data [172]. These algorithms are mostly based on correlations between the various components (*variables*) of the data. These correlations can be affected by the application of data-privatization mechanisms aimed at protecting the privacy of data providers. However, protecting data privacy is a legal obligation in Europe and many other countries around the world. In response to this necessity, numerous privatization methods have been developed to maximize the trade-off between a good level of data privacy and utility.

In general, the addition of noise tends to reduce the utility of the information that can be extracted from the data. Many privatization approaches and denoising techniques have been optimized for the summary statistics of the individual variables in the data, such as average values. However, notions of utility also depend on the correlation between the various components of the data, especially in the case of causal discovery. Some approaches to cope with this problem have been proposed in the global DP setting when the full unobfuscated data set is available. For example, the collected data may be synthesized using generative algorithms such as GAN [173] or Bayesian Networks [174]. However, little or no instances of relation-preserving local DP mechanisms are known for causal discovery. In the local setting, the data are already obfuscated before they get to the central server, and therefore the methods used in global DP are not applicable.

In this chapter, we experimentally assess the impact of state-of-the-art LDP and d-privacy mechanisms on the structural accuracy of causal discovery from the data. More precisely, as the LDP representative, we consider the k-Ary Randomized Response (k-RR, Section 9.2.1) [175]. As

the local d-privacy representative, we considered the Geometric mechanism (Section 9.2.1). We conduct extensive experiments on both real and synthetic data sets and evaluate their impact on 9 causal discovery algorithms, including constraint-based, score-based, and causal asymmetry-based methods.

9.2 Related work

Causal discovery with DP is an emerging research area that aims to combine the benefits of both identification of causal relationships among variables and privacy-preserving data analysis. The goal is to discover causal relationships between variables while preserving the privacy of sensitive data. An approach explored in the literature for differentially private causal discovery was to incorporate DP mechanisms directly into existing causal discovery algorithms [176–179]. These algorithms introduce controlled noise during the causal learning process to ensure privacy protection.

However, these existing differentially private causal discovery algorithms assume the centralized DP model, which requires collecting users' original data. The approach adopted in this thesis is to leverage the concept of local DP [36, 37] for causal discovery (respectively local d-privacy [41]). In recent years, several works have been done in the local DP setting (e.g., see [37–40, 175, 180–184] and references within), and applying them to causal discovery involves sanitizing the data at the individual level. Parallel to our work, [185] studies a class of corruptions, such as measurement error, missing values, discretization, and differential privacy in the US Census. However, their goal is to learn a causal parameter (average treatment effect) from corrupted data, and they conduct experiments only in an aggregated setting. Similarly, [186] offers causal inferential methodologies for analyzing locally differentially private data. [67] experiment with causal discovery with a small amount of noise added to the data. However, the noise is not produced by the privatization mechanism. These goals differ from our work; they investigate the effect of noise in causal effect estimation of a treatment (or intervention) when randomized experiments are impossible to conduct, thus statistical theory is needed. Our work solely focuses on causal discovery, that is, the inference of causal *relations*, causal *directions* among a set of variables (i.e., "how the change in X influences Y?" versus "is X the cause of Y?"). To the author's knowledge, this is the first work that thoroughly explores and analyzes the impact of locally differentially private mechanisms on causal discovery.

9.2.1 Privacy Mechanisms

In this section, we describe the various discrete multidimensional mechanisms used in this chapter. Visually, Figure C.1 in Appendix C.1 shows the four mechanisms applied to a single point in a 4D space with shape $(2, 5, 5, 5)$, denoting the number of categories or bins per dimension.

k-ary Randomized Response (k-RR)

Randomized response (RR) was proposed in [187] with the aim of providing "plausible deniability" to individuals responding to embarrassing (binary) questions in a survey. [175] generalized RR to domains of arbitrary size k (with $k \geq 2$), and proposed the so-called k-RR mechanism, which is one classical technique for achieving LDP in categorical / discrete data. Given a data domain V, and the privacy parameter ϵ, let $k = |V|$ and $p := \frac{e^\epsilon}{k-1+e^\epsilon} \in (0,1)$. For each $v \in V$, let $\eta_{\neq v} \in V$ be a uniform random variable (i.e., exogenous noise with uniform distribution) over $V \setminus \{v\}$. We let k-RR : $V \to V$ be the random variable given by:

$$k\text{-RR}(v;\epsilon) := \begin{cases} v, & \text{with probability } p \\ \eta_{\neq v}, & \text{with probability } 1-p \ . \end{cases}$$

This mechanism satisfies ϵ-LDP [175], because $\frac{p}{q} = e^\epsilon$, where $q := \frac{(1-p)}{(k-1)}$. When collecting data in practice, one is often interested in multiple attributes of a population, i.e., *multidimensional data*. We assume that there are d attributes with domains $A_1, A_2, \ldots, A_d$, where each A_i is a discrete set of finite size $k_i = |A_i|$. Each data provider u_j for $j \in \{1, 2, ..., n\}$ contributes to the data set with a tuple (record) $\mathbf{v}^{(j)} = (v_1^{(j)}, v_2^{(j)}, ..., v_d^{(j)})$, where $v_i^{(j)}$ represents the value of the attribute A_i. We now describe the two main known methods for applying k-RR on multidimensional data [182, 183, 188].

k-RR component-wise (k-RR C-wise). This is a naive approach that applies k-RR independently on each attribute. More precisely, k-RR C-wise splits the privacy budget ϵ among the d attributes uniformly or proportionally to their size, and reports each attribute in A_i using k_i-RR parameterized with ϵ_i-LDP, where $\epsilon_i = \epsilon \cdot \frac{k_i}{k_1+k_2+\ldots+k_d}$.

k-RR Combined (k-RR Comb). This mechanism considers the Cartesian product $A_1 \times A_2 \times \ldots \times A_d$ as a single attribute and sanitizes it using k-RR parameterized with ϵ-LDP, where $k = k_1 \cdot k_2 \cdot \ldots \cdot k_d$.

Bounded geometric mechanism

The geometric mechanism is the discrete analogue of the Laplace mechanism. The output Y is related to the input X by the formula:

$$\mathbb{P}[Y = y | X = x] = p_{\max} \exp(-\epsilon |y - x|) \tag{9.1}$$

for some parameters ϵ that represents the level of privacy. $p_{\max}$ is a normalization factor, that is, it is chosen so that $\sum_y \mathbb{P}[Y = y|X = x] = 1$. This formula is valid in 1D, in which $|\cdot|$ denotes the absolute value, as well as in multidimensional Euclidean space, in which x and y are discrete vectors and $|\cdot|$ denotes the Euclidean norm, or any other p-norm chosen in advance (see

Figure C.2 for a comparison). From the definition of the geometric mechanism, it is immediate that it satisfies local d-privacy with privacy parameter ϵ, where the metric d is the chosen p-norm based distance.

In this chapter, we are interested in bounding the geometric mechanism so that the output domain equals the input domain, as in k-RR. There are three natural ways to do it, namely (1) clipping, (2) replacing samples that are out of the box with uniform noise, and (3) resampling whenever a sample is out of the box. Let us review them in more detail.

Method (1), clipping, consists of replacing all output values that lie outside the box with the closest values that lie inside the box, that is, with the maximum or minimum values of the domain in the 1D case. In this case, the two extremes of the box may increase their probabilities excessively, and the property that the output y with maximum probability is always $y = x$ can be lost, especially when the input x is close to the border. In Method (2), whenever the output y is outside the box, it is replaced with a uniform sample from the box. In terms of the probability distribution of the mechanism, this method crops it from the background (two tails in the 1D case), and rescales the cropped distribution by adding a constant. This addition results in combinations of exponential terms with additive constants, which unnecessarily adds complexity to the formulas and distorts the exponential shape and its decay properties. Instead, in Method (3), which corresponds to sampling as many times as necessary until the output is inside the box, the cropped distribution is simply multiplied by a constant. This preserves the main shape of the distribution, while also keeping the formulas relatively simple. For this reason, we prefer method (3) over the other two.

Notice that the bounding is not symmetric, except for the input in the center of the box. This means, that we should have different values of $p_{\max}$ or ϵ for different values of x so that the bounded summation is 1 on all x. As it will be justified in Section 9.3, we opt for fixing $p_{\max}$, so the formula that characterizes the bounded geometric mechanism becomes:

$$\mathbb{P}[Y = y|X = x] = p_{\max} \exp(-\epsilon_x|y - x|)$$

where both x and y are constrained to a fixed bounded discrete set, and ϵ_x are chosen so that $\sum_y \mathbb{P}[Y = y|X = x] = 1$. These values always exist (assuming $p_{\max} \geq 1/k$), and we provide an algorithm to find them.

The computation of ϵ_x for every x is not possible symbolically through a formula. It is required that $\sum_y \mathbb{P}[Y = y|X = x] = 1$, or equivalently, $\sum_y \exp(-\epsilon_x|y - x|) = \frac{1}{p_{\max}}$, where both x and y are constrained to a fixed bounded discrete set. In the 1D case, the domain is a set of k contiguous integers and for the smallest value of x, only one tail of the geometric distribution intersects the domain, which allows us to write $\frac{1}{p_{\max}} = \sum_y \exp(-\epsilon_x|y - x|) = \sum_{\delta=0}^{k-1} \exp(-\epsilon_x k) = \frac{1-\exp(-k\epsilon_x)}{1-\exp(\epsilon_x)}$. However, there is no analytical solution for ϵ_x from this formula. Moreover, for the remaining values of x, the expression becomes more complex, as an additional term is added for the second tail, and even more so for the multidimensional case.

However, the computation of each ϵ_x can be carried out numerically exploiting the fact that $\sum_y \exp(-\epsilon_x|y - x|)$ is decreasing on ϵ_x. At one extreme, if $\epsilon_x \to 0$, the sum approaches k, and at

the other, if $\epsilon_x \to \infty$, the sum approaches 1. This implies, first, that there is a unique point ϵ_x for which this function crosses the threshold $\frac{1}{p_{\max}}$, and more importantly, that we can use a binary search to compute ϵ_x. In the multivariate domain, the summations still satisfy the monotonicity property. Therefore, this method can be used to implement the multidimensional geometric distribution. Similar to k-RR, we compare two versions of the geometric mechanisms, namely, Geo Comb and Geo C-Wise.

9.3 Tuning the Level of Privacy

The parameter ϵ in LDP does not have the same meaning as the ϵ in d-privacy, i.e., they represent different levels of privacy. In order to compare the mechanisms of these two families, we need to tune the respective ϵ's so as to represent the same level of privacy. To avoid confusion for the readers that know the standard notion of DP, and are not so familiar with LDP, it is important to remind the reader that the standard notion for privacy in the local framework is not the same as in the central one: In central DP, the challenge for an attacker is to distinguish between two adjacent data sets, i.e., data sets that differ for presence or absence of one record. In other words, the attacker wants to infer whether or not a certain record is in the data set. In LDP, in contrast, the aim of the attacker is to infer the true value of the individual data provider.

To measure the level of privacy, therefore, we consider the probability that an attacker has to infer the true value from the reported value. Naturally, the attacker will bet on the value that has the maximum posterior probability, given the obfuscated value [189, 190]. We note that this measure of privacy is directly related to the notion of *advantage of an attacker* in security, and to the notion used to assess the vulnerability of the training set in ML.

In both k-RR and d-privacy, the value that has the highest probability to be reported is the true value itself, hence the level of privacy provided by these mechanisms (assuming a uniform prior) is the probability to report the true value. Specifically, the level of privacy provided by k-RR with parameter ϵ is:

$$Priv_{k\text{-}RR}(\epsilon) \;:=\; \frac{e^{\epsilon}}{k-1+e^{\epsilon}}\;.$$

whereas, for a Geometric with parameter ϵ', the level of privacy is:

$$Priv_{Geo}(\epsilon') \;:=\; \mathbb{P}_{\max}\cdot e^{\epsilon'\cdot 0} \;=\; \mathbb{P}_{\max}\;.$$

where $p_{\max}$ is the normalization factor used in the definition of the geometric mechanism (Equation (9.1)). Tuning the parameters of k-RR and L to provide the same level of privacy means adjusting the above ϵ and ϵ' so that $Priv_{k\text{-}RR}(\epsilon)$ and $Priv_{Geo}(\epsilon')$ give the same result.

9.4 Experimental Results

In this section, we empirically assess how locally private mechanisms impact causal discovery. We evaluate the performance of 9 causal discovery algorithms in multidimensional, two-dimensional, real and synthetic data sets obfuscated using the various mechanisms described in Section 9.2.1.

We start by applying the algorithms to discretized non-obfuscated data. Then we select the algorithms that performed best at a particular data set and apply them to the obfuscated versions of this data set. We measure the effect of each privatization method on the algorithms by comparing the Structural Hamming Distance (SHD) score and the F1 score or Accuracy on the non-obfuscated and obfuscated data. We use the Benchpress causal discovery benchmarking framework [191] to generate synthetic data and run causal discovery algorithms for multidimensional experiments. As (L)DP mechanisms are randomized, we report average results over 5 runs. Due to space constraints, we have included all of our additional experiments in Appendix C.2.

9.4.1 Data Sets

We use real benchmark and synthetic data sets for the experiments. The details can be found in Table 9.1.

Name	Type	Nodes	Bins	Size	Origin
Sachs	real	11	10	902	[192]
Human Stature	real	3	10	898	[193]
Synth10	synthetic	10	10	5000	random DAG, IID, Linear, Gaussian
Synth5	synthetic	5	5	50000	random DAG, IID, Linear Gaussian
CEP	real	2	2-100	94-16382	[67]

Table 9.1: Data sets used for causal discovery. For CEP the number of bins was determined by $min(u, 100, u * 0.1)$, where u denotes the number of distinct values.

The Sachs data set measures the expression levels of various proteins and phospholipids within human cells. It was originally generated by [192]. The data set consists of 11 variables and 902 samples. Sachs is a popular benchmarking data set in causal discovery because of availability of the ground-truth causal structure.

Human Stature data set id a classic historical data set collected by the statistician Francis Galton and first used for regression analysis [194]. Later it has been re-used as one of the benchmark data sets for causal discovery. The data set consists of four variables: father height, mother height, gender, and child height, and has 898 samples. We remove the binary gender variable for our experiments. We do it because when applied to binary data, geometric noise becomes equivalent to k-RR method.

Synth10 and Synth5 are synthetic data sets with 10 and 5 nodes, respectively. The background structure DAG is generated randomly using the benchpress framework [191]. We specify the number of nodes and the maximum number of parents for each node. The data are generated using a generation process compatible with the underlying structure of the DAG.

CEP data set [67] is a collection of data sets of causal pairs in the real world, such as, for example, altitude and temperature. The collection consists of 99 data sets of varying sizes, however, it is often referred to as a single data set for causal discovery benchmarking.

9.4.2 Causal Discovery Algorithms

We apply constraint-based and score-based causal discovery algorithms for multidimensional data. We select several well-known algorithms that can run on discretized data. For pairwise data sets, we apply algorithms that are capable of identifying the causal direction for two variables. We test the performance of the discrete and continuous data-specific versions of the algorithms, as well as various parameter values. The details can be found in Table 9.2 and in the Preliminaries 5.5.1.

We have used two libraries for implementation: we used the Benchpress [191] package for the PC, FCI, FGES, Iterative MCMC and MMHC causal discovery algorithms and metrics, for the RECI, IGCI, CDS and ANM methods we used the Causal Discovery Toolbox [195].

Algorithm	CI Test/Score	Parameter
PC ([114])	Gaussian, Chi-square	Alpha (0.001,0.05, 0.1)
FCI ([115])	Fisher-Z, Chi-square	Alpha (0.01,0.05,0.1)
FGES ([116])	BIC	Penalty discount (0.75,0.8,1,1.5)
Iterative MCMC ([119])	BGe	Alpha (0.001,0.01,0.1)
MMHC ([120])	BDe	Alpha (0.01,0.05, 0.1)
RECI ([121])	MSE	
IGCI ([122])	sp1	
CDS ([196])	std. dev.	Forced Decision
ANM ([123])	HSIC	

Table 9.2: The structure learning algorithms.

9.4.3 Discretization

In order to apply the discrete mechanisms of interest to our dataset, it was necessary to discretize the original continuous data. Discretization is a critical step in the process, as it plays a pivotal role in the subsequent data analysis. There are several approaches to discretizing data, each with varying effects on the quality of the results. Some of these methods yield higher average precision, up to the highest possible [197], but rely on knowledge of properties about the underlying data distribution, such as quantiles or an estimation of the density function. However, in situations where the underlying data is sensitive and private, revealing such properties can risk privacy breaches, so it is safer to assume that they are unknown. To address this challenge, we opted for the simplest method of discretization, namely uniform bins within a fixed range. In practice, this fixed range corresponds to estimations of the minimum and maximum values of the population.

The only parameter we can freely choose in this process is, therefore, the number of bins and it should be chosen taking into consideration that more bins imply more accurate information being revealed. Moreover, the number of dimensions of the data, which corresponds to the number of columns in the data set, also plays a role in the choice of the number of bins, as it increases exponentially the total number of bins. We chose between 5 and 10 bins for datasets with 3 or more dimensions, and for the CEP data set, which has two dimensions but contains several different data sets, we applied a dynamic number of bins (see in 9.1). Some of these datasets were already discretized (e.g. had only 2 distinct values), and some had continuous

data. We determined the number of bins by $min(u, 100, u * 0.1)$, where u denotes the number of distinct values in a given data set.

9.4.4 Evaluation metrics

For the data sets with more than two-dimensions we used structural hamming distance (SHD) to measure the difference between the ground truth adjacency matrix and the output of the causal discovery algorithm. It assigns a distance of 1 for every missing, redundant or reversed edge in the graph. Intuitively, SHD provides a number of edges that are need to be added, removed and re-directed to make the two graphs identical. We have also calculated the F1 score, that combines the precision and recall of a model, and is used to evaluate the recovery of the skeleton of the DAG.

In case of the CEP data set we have applied the same method as in [67]. Forced-decision: given a sample of a pair (X, Y) the methods *must* decide on a causal direction. Then, we evaluate the weighted [1] accuracy of the decisions. We also calculate the confidence intervals assuming a binomial distribution using the method by [198].

9.4.5 Results on Multidimensional Data

We report the results for the algorithms that performed the best on the discretized, but not obfuscated data. PC algorithm performed the best on most of the data sets. Iterative MCMC algorithm was performing better on the data sets with 10 or more nodes. Both data sets with 10 or more nodes show that causal discovery algorithms in general perform better under geometric privatization methods rather than k-RR. For Sachs data set (Figure 9.1) PC and GES algorithms perform almost the same on Geo C-wise and Geo Comb. The performance on geometric mechanism is very close to the performance on the original data without the noise. For Synth10 data (Figure 9.3) the performance on data obfuscated with geometric mechanisms with $p_{\max} = 0.5$ outperform the results on the original data. However, this result can as well be accidental. For Synth10 data we also observe a slightly better performance when Geo Comb is applied as compared to Geo C-wise. Performance is better with k-RR C-wise privatization than with k-RR Comb privatization on the Sachs and Synth10 data sets. For smaller multidimensional data sets (Figures 9.2 and 9.4) the variation of the performance is too large to draw reliable conclusions. This is probably due to the high influence of chance on recovering the data structure when the true graph is small. However, we still observe a slight advantage in applying geometric mechanisms to Synth5 and Human Stature data sets. We can also observe slightly better SHD results with K-RR C-wise privatization than with K-RR Comb privatization on Synth5 and Human Stature data sets.

In our additional experiments in Appendix C.2.1, we observe similar results when measuring the F1 score for the causal discovery of an undirected graphs (Figures C.3, C.10, C.17, C.24).

[1] Not all pairs can be considered as independent. Weights' list was acquired from the authors' website.

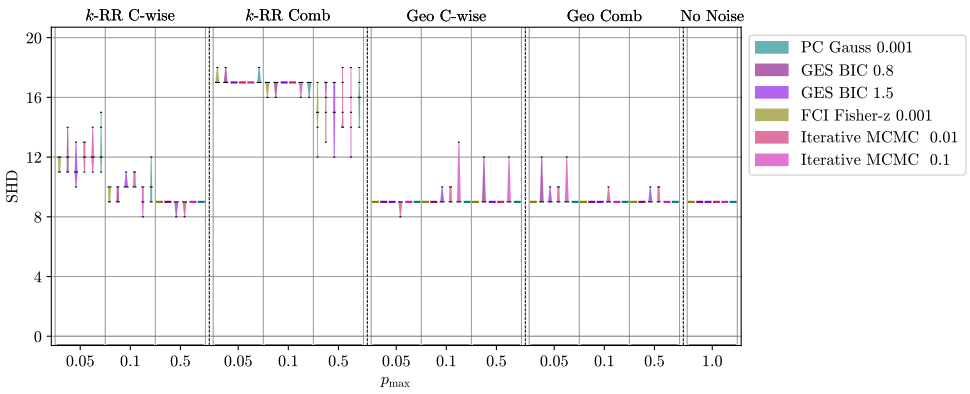


Figure 9.1: Sachs data, SHD. The results for PC algorithm with Gaussian CI test and alpha value 0.001; GES algorithm with BIC score and penalty discount values 0.8 and 1.5; FCI algorithm with Fisher-z CI test and alpha values 0.001; Iterative MCMC algorithm with BGe score and alpha values 0.01 and 0.1. The width of each bar varies for different values on the y-axis proportionally to the number of samples attaining that value.

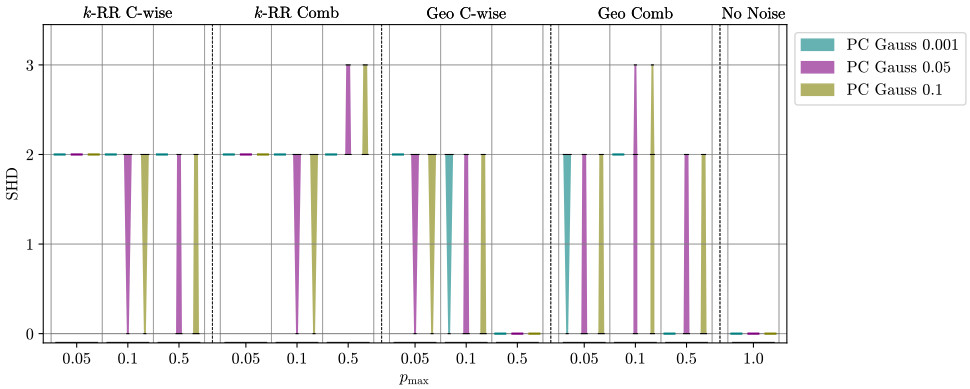


Figure 9.2: Human Stature data, SHD. Results for PC algorithm with Gaussian CI test and alpha values 0.001, 0.05 and 0.1.The width of each bar varies for different values on the y-axis proportionally to the number of samples attaining that value.

9.4.6 Results on Two-dimensional Data

We report the results of all causal discovery algorithms applied for the CEP data set. In Figure 9.5, we show the results before (“No Noise") and after privatization. It is evident that, similar to previous experiments, the geometric mechanism consistently outperforms k-RR, with notable improvements, especially in the case of RECI, where the accuracy surpasses the baseline. We hypothesize that this phenomenon could be attributed to the potential data augmentation properties of noise addition, although further research is required to confirm this. The CDS algorithm performs similarly after privatization, except when applying the k-RR Comb mechanism. But k-RR Comb generally has the poorest performance (also with Sachs and HS datasets), we think this is due to the available small sample size, and the mechanism is affected by the curse of dimensionality. IGCI’s accuracy drops by approximately 15-20%, however, this is not surprising. The IGCI model’s practical applicability is limited to causal relations with sufficiently small noise and its drop in performance has already been shown by [67], where the added noise was very little (much smaller than in our experiments). ANM exhibited unsatisfactory performance even before noise introduction, and its performance deteriorated further (sometimes falling below

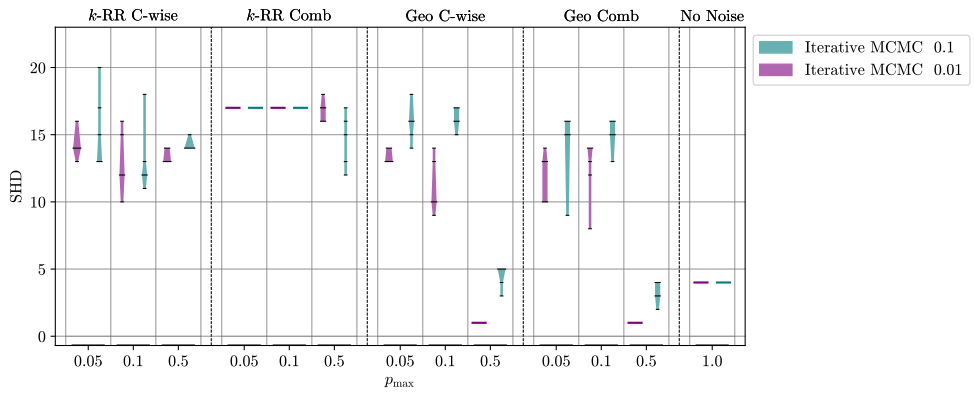


Figure 9.3: Synthetic data, 10 nodes, SHD. The results for Iterative MCMC algorithm with BGe score and alpha values 0.01 and 0.1. The width of each bar varies for different values on the y-axis proportionally to the number of samples attaining that value.

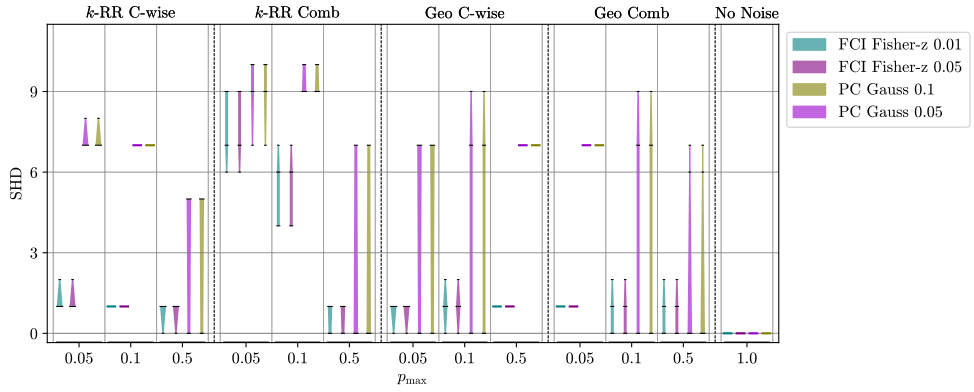


Figure 9.4: Synthetic data, 5 nodes, SHD. The results for FCI algorithm with Fisher-z CI test, alpha values 0.01 and 0.05; PC algorithm with Gaussian CI test, alpha values 0.1 and 0.05. The width of each bar varies for different values on the y-axis proportionally to the number of samples attaining that value.

chance levels) after privatization.

9.5 Discussion

Our results consistently demonstrate that **geometric privatization methods (both component-wise and combined) exhibit higher accuracy in terms of SHD compared to k-RR methods (both component-wise and combined).** In case of geometric noise, the algorithms do not appear to perform much worse as the noise increases. This can be expected because this privatization method is not disruptive of the correlations in the data. It would be an interesting extension to also evaluate its effect on the model parameters. On the other hand k-RR noise interrupts with the data structure, and more noise results in worse performance of the causal discovery algorithms. We observe similar results when measuring the performance of causal discovery algorithms with the F1 score.

We observe some dependence between the higher parameter alpha (PC) or penalty discount (GES) parameters and better F1 scores on the noisy data in the experiments on multidimensional data. Higher parameter values result in sparser graphs and help avoid spurious edges in the

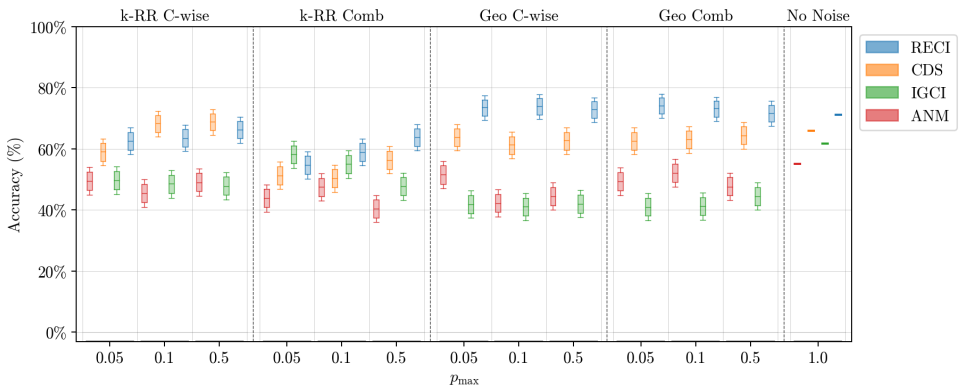


Figure 9.5: CEP data set with 2 nodes, weighted accuracy. Box whiskers are at 95%, body is at 80% confidence.

graphs. We observe that algorithms that are less accurate on the original data are also less sensitive to data privatization. More precisely, when applied to obfuscated data, their performance drops less compared to the baseline on the original data (the detailed results can be found in an Appendix C.2). However, the algorithms which are best on the original data still provide the best overall results under geometric noise (despite being more sensitive to k-RR noise).

Part V

Causality

10

On the Need and Applicability of Causality for Fair Machine Learning

10.1 Introduction

Accurate measurement of discrimination is important for evaluating the data or algorithm and for advising methods for achieving fairness. Recently, the domain of AI fairness has seen an increase in the use of statistical causality methods to evaluate and mitigate discrimination in data and algorithmic decisions.

This article consolidates the statistical and legal arguments for using causality in fair AI as well as practical challenges. We argue that causality is needed to appropriately address the problem of fairness in ML based automated decision systems. We summarize the benefits of using causality in three arguments, namely, (1) reliably measuring discrimination, (2) mediation analysis, and (3) establishing causal evidence in legal practice. Compared to existing work, the latter argument can be seen as the first attempt to connect causality in fair AI with the European AI legislation.

Tackling the problem of fairness from a causal perspective is plagued by practical obstacles that hinder its use in real scenarios. This includes the existence of several constraining assumptions that need to be satisfied and the availability of the causal graph. The last part of the chapter describes the different assumptions and discusses their implications in the specific context of ML fairness.

Measurement of discrimination without taking into consideration the causal structure underlying the relationships between variables may lead to misleading conclusions. That is, a biased estimation of discrimination. In extreme cases, such as Simpson's paradox, the bias may lead to

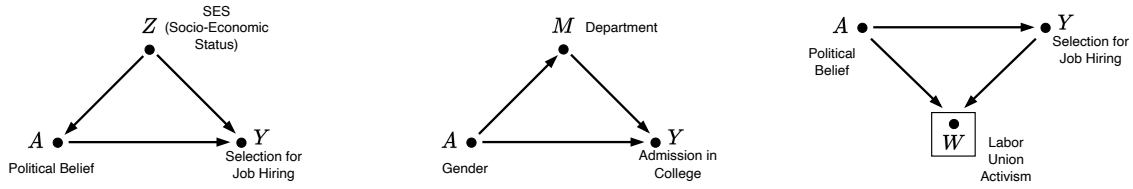
reversing the conclusions (e.g. the biased estimation indicates a positive discrimination, while the unbiased estimation is actually a negative discrimination). Figures 10.1(a)-10.1(c) show the three basic causal structures that can lead to statistical anomalies, and consequently make common statistical metrics of fairness unreliable.

10.1.1 Related Work

Despite the increase in specific applications of the causal approach, the general discussion of the benefits and challenges of adaptation of causal frameworks to fair machine learning is very limited. Most articles provide specific solutions to causal fairness problems, give general arguments for avoiding spurious correlations, and make strong assumptions, for example about the availability of causal structure, without further consideration [108, 111, 199–201]. Loftus et al. [202] summarize the advantages of using causality in Fair machine learning. Several more recent studies warn about the dangers of using counterfactual models either due to their sensitivity to unmeasured confounding or incompatibility with social reality [203, 204]. [205] discusses the assumption of ignorability in the context of fairness in ML. [18, 206] lay out the arguments for causality in a legal context and do not clearly link the judicial process with causal mediation analysis. Researchers in [207] discuss the compatibility of the notions of fairness of ML with the legal applications. However, they do not focus on causal fairness notions or European law.

10.2 Reliably measuring discrimination

In this section, we revisit basic causal structures such as confounder, collider, and mediator by placing the fairness context. We discuss the importance of collider or confounder bias in measuring fairness. In addition, we consider mediation analysis as a tool for better understanding the mechanism behind disparity.



((a)) Confounder structure. ((b)) Mediator structure. ((c)) Collider structure.

Figure 10.1: Basic causal structures in fairness context.

10.2.1 Confounder structure

The first situation where ignoring the causal structure of the data may lead to an unreliable estimation of discrimination is due to a failure to consider a confounder variable. Consider the hypothetical example in Figure 10.1(a) of an automated system for selecting candidates for job positions. Assume that the system takes as input two characteristics, namely, the socioeconomic status (SES) denoted as Z and the political belief of the candidate A. The outcome Y is whether

the candidate is selected for the next stage of hiring (or the probability that the candidate is selected). The outcome Y is influenced by the SES (a better SES makes it possible for candidates to attend more reputable academic institutions and enroll in costly trainings). Both variables can be either binary (Z could be rich or poor, while A could be liberal or conservative) or continuous (how rich/poor a candidate is for Z and the degree of conservativeness of the candidate for A). The political belief A of a candidate can be influenced by several variables, but in this example, assume that it is only influenced by the SES of the candidate. Finally, assume that the automated decision system is suspected to be biased by the political belief of candidates. That is, it is claimed that the system will more likely select candidates with a particular political belief.

A simple approach to check the fairness of automated selection Y with respect to the sensitive attribute A is to contrast the conditional probabilities: $\mathbb{P}(Y = 1 \mid A = 0)$ and $\mathbb{P}(Y = 1 \mid A = 1)$, corresponding to statistical disparity, which quantifies the disparity in the selection rates between both types of candidates (conservatives and liberals). However, such an estimation of discrimination is biased due to the confounding path through Z. As the variable Z causes both the sensitive variable A and the outcome Y, it creates a correlation between A and Y which is not causal. In other words, high SES (rich) candidates tend to have a more conservative political belief and at the same time more chances to be selected for the job (better academic institutions and training), which creates the following correlation in the data: employers will have more candidates with conservative political beliefs, and hence less candidates with liberal political beliefs. This correlation is due to the confounder Z and should not count as discrimination. Most statistical notions of fairness (equal opportunity, predictive parity, etc.) are not suitable to measure discrimination in the presence of such statistical anomaly.

10.2.2 Mediator structure

The second situation where not accounting for the causal structure behind the data may lead to unreliable estimation of discrimination involves the presence of one or several mediator variables. The problem emerges from whether to consider discrimination through a mediator variable as justifiable/acceptable or not. Similarly to confounding structure, a mediator variable may lead to Simpson's paradox. A famous example of Simpson's paradox caused by a mediator structure is the gender bias in 1973 Berkley admission [20, 208]. Figure 10.1(b) shows the causal graph underlying the data, where the sensitive variable (A) is gender, the outcome (Y) is admission for Berkley graduate studies, and a single mediator variable (M) representing the department for which a candidate applied. In 1973, 44% of male applicants were admitted against only 34% of female applicants. Although this seems like a bias against female candidates, when the same data were analyzed by department, acceptance rates were approximately the same. In a simple mediator structure, there are two possible paths from A to Y: a direct path $A \to Y$ and an indirect path $A \to M \to Y$. Comparing the global admission rates of male and female candidates corresponds to considering both paths when measuring discrimination. Whereas, comparing the admission rates per department corresponds to considering only the direct path $A \to Y$. Hence, whether or not to consider mediator paths when measuring discrimination may lead to contradictory conclusions, such as in Simpson's paradox.

10.2.3 Collider structure

A biased estimation of discrimination may be due to the presence of common effect (collider) variable and a data generation process implicitly conditioning on that variable. Using the same hypothetical example of job selection, consider the causal graph in Figure 10.1(c). A and Y are the same as in the previous example. Assume that data for training the automated decision system is collected from different sources, but mainly from labor union records. Assume also that the variable W representing the labor union activism of the candidate is caused by both A and Y. On the one hand, political belief A influences whether a candidate is an active member of labor union (individuals with liberal political beliefs are more likely to enroll in labor unions). On the other hand, if a candidate is selected/hired, then there are higher chances that she becomes a member of labor union and consequently that her case is recorded in the labor union records. Consistent with previous work, a box around a variable (W) indicates that the data is generated by implicitly conditioning on that variable.

Again, the simple approach of contrasting the selection rates between both types of candidates (conservatives and liberals) leads to a biased estimation of discrimination due to the colliding path through W. Intuitively, an individual has a record in the collected data either because she has liberal political beliefs or because she is selected for the job. Individuals who happen to have liberal political beliefs and at the same time selected for the job are still present in the data; however, conditioning on labor union activism creates a correlation between A and Y which is not causal: data coming from labor union records includes fewer liberal candidates which are selected for the job than conservative candidates. Again, this is discrimination against candidates with liberal political beliefs. Such correlation is due to the collider structure and should not count as discrimination.

10.3 Mediation Analysis

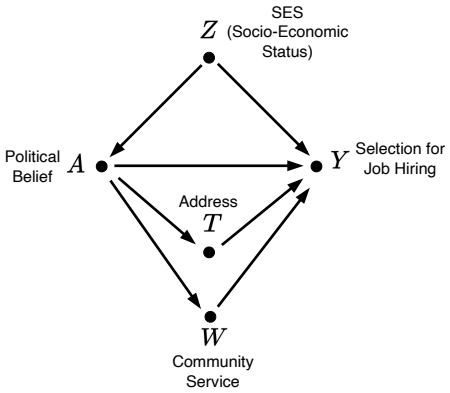


Figure 10.2: Causal graph with two mediated paths.

In presence of one or several mediator variables, it is useful to know how much discrimination is direct and how much is mediated. More precisely, how much discrimination is conveyed through each mediator variable. Mediation analysis is about distinguishing the different paths through

which discrimination is going through and the portion of discrimination conveyed through each path. Consider another variant of the job-hiring example in Figure 10.2 with two mediator variables, address (T) and community service (W). There are in total four different paths from the sensitive variable A (political belief) to the outcome variable Y (job hiring):

- $A \leftarrow Z \rightarrow Y$: confounding path
- $A \rightarrow Y$: direct path
- $A \rightarrow T \rightarrow Y$: indirect path through T
- $A \rightarrow W \rightarrow Y$: indirect path through W.

The first confounding path is non-causal, and hence any effect slipping away through it should not be considered when estimating discrimination. As described in Section 10.2.1, this spurious effect is due to how the data is generated / collected and consequently should not count as actual discrimination. The direct path is present whenever there is an edge between A and Y. The effect through $A \rightarrow Y$ is always discriminatory, that is, it can never be justified and considered acceptable discrimination. The two remaining paths are indirect paths going through mediator variables. Discrimination through an indirect path can or cannot be justified depending on the nature of the mediator variable. For example, in the example of hiring a job in Figure 10.2, T (home address) is a mediator variable because, on one hand, having a certain political inclination may indicate where a candidate is living and, on the other hand, the job hiring decision may depend on the home address of a candidate. W (community service) is another mediator variable because, on the one hand, the political views of a candidate can influence how much involved she can be in community service, and on the other hand, the community service record is a good indicator on how suitable she will be for a given position. Discrimination on the path $A \rightarrow W \rightarrow Y$ can be acceptable as an employer can justify a disparity between candidates with different political beliefs by their community service records. However, discrimination through the path $A \rightarrow T \rightarrow Y$ is typically not acceptable because an employer cannot justify discrimination on the basis of the addresses of candidates. T is called a proxy variable, whereas W is called an explanation variable[1].

Causality, through the concepts of intervention and counterfactual, provides the tools required to distinguish between discrimination conveyed through different paths. Intervening on A, blocks all paths from an incoming edge to A which include all confounding paths between A and Y. Discrimination through all causal paths is captured by the average causal effect (ACE):

$$ACE(Y, A) = \mathbb{P}(Y = y^+|do(A = 1)) - \mathbb{P}(Y = y^+|do(A = 0)) \tag{10.1}$$

where $Y = y^+$ is a positive decision and $A = 1, A = 0$ are the values of the sensitive attribute. In Figure 10.2, ACE expression captures discrimination through all paths, except $A \leftarrow Z \rightarrow Y$. For simplicity of notation, we represent $Y = y^+$ simply as y^+, $A = 1$ (resp. $A = 0$) as a_1 (resp. a_0) and

[1] In presence of a single path with a sequence of two or more mediators, the existence of at least one explaining variable among the mediators makes discrimination through that path justifiable and hence acceptable.

the $do()$ operator with subscription. Therefore, the right-hand size of Equation 10.1 becomes simply $\mathbb{P}(y^+_{a_1}) - \mathbb{P}(y^+_{a_0})$.

To distinguish the direct discrimination from indirect discrimination, two expressions can be used, namely natural direct effect (NDE) and natural indirect effect (NIE) [107]:

$$NDE(Y, A) = \mathbb{P}(y^+_{a_1, \mathbf{Z}_{a_0}}) - \mathbb{P}(y^+_{a_0}) \tag{10.2}$$

where $\mathbf{Z}$ is the set of all mediator variables and $\mathbb{P}(y^+_{a_1, \mathbf{Z}_{a_0}})$ is the probability of a counterfactual situation where $Y = y^+$ had A been 1 and had $\mathbf{Z}$ been the value it would naturally take if $A = 0$. Intuitively, $\mathbb{P}(y^+_{a_1, \mathbf{Z}_{a_0}})$ is considered counterfactual because it corresponds to a candidate who is conservative ($A = 1$) on the direct path $A \rightarrow Y$ but liberal $A = 0$ on all indirect paths. NIE has a similar form but A values are reversed in the counterfactual expression:

$$NIE(Y, A) = \mathbb{P}(y^+_{a_0, \mathbf{Z}_{a_1}}) - \mathbb{P}(y^+_{a_0}) \tag{10.3}$$

Finally, distinguishing the discrimination conveyed through specific indirect paths is possible through path-specific effect (PSE) [107, 108]:

$$PSE(Y, A, \pi) = \mathbb{P}(y^+_{a_1|\pi, a_0|\bar{\pi}}) - \mathbb{P}(y^+_{a_0}) \tag{10.4}$$

where π is set of the variables on the path of interest, $\bar{\pi}$ is the set of variables not in π and $\mathbb{P}(y^+_{a_1|\pi, a_0|\bar{\pi}})$ is the counterfactual probability of $Y = y^+$ had A been 1 on the paths π and 0 on the remaining paths $\bar{\pi}$.

10.4 Uncovering causality through legal evidence: the regulatory approach in the European Union

The method of mediator structure in uncovering causation, as discussed in the previous section, is certainly a useful model for the proof of causality in judicial instances dealing with algorithmic discrimination. However, the question is whether procedural law, namely in the European Union (EU), is designed to support such an analysis. As a preliminary observation, we should stress that in law, the expression 'causal fairness' generally refers to the procedural conditions under which instances of fairness (or unfairness, for that matter) are causally represented. With this in mind, in this section, we will focus on two important and interrelated issues: evidence and procedural fairness [2]. In the eye of the law, causality is a question of fact, calling for legally established discovery procedures - and corresponding reasoning models - meant to yield accurate causal representations, i.e., allow for causality proper to be singled out from a myriad of correlations (positive associations between candidate-causes and a harm suffered) [209]. However, in adjudicatory contexts, causality is proven for the purpose of fairness, typically

[2] In the EU, the Independent High Level Expert Group on AI, set up by the European Commission, defined procedural fairness as "entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them." See HLEG, Ethics Guidelines for trustworthy AI, available at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, at 13.

compensation as a "fair" outcome to the suffering of harm. In fact, legal systems committed to the rule of law [3] share a commitment to procedural fairness, the normative creed being that only fairly designed procedures can be conducive to fair outcomes. In contemporary systems of evidence and judicial remedies, including those in EU law, the fair procedures/outcomes parallelism is epitomized in the fair trial safeguards - procedural entitlements meant to uphold a level of basic equality (or procedural parity) and effectiveness in the ways in which litigants participate in a dispute resolution [210]. This equality not only applies to the litigants ability to access judicial remedies, but also to their ability to access and give evidence, the idea being that one party should in no way be advantaged or disadvantaged over the other, in terms of their access to the facts needed to make their views known (usually, before a court). In short, "casual fairness" in law calls for accurate - or at least, plausible [211] - evidence of causality, presented in conditions of procedural fairness.

Proof of causality in connection to algorithmic discrimination has profoundly upset these longstanding legal postulates. From a procedural fairness perspective, a major thorny issue has been that AI's relative or total opacity makes AI systems' decisional processes inscrutable, obstructing the victims' ability to properly establish and argue causation. One of the topical examples in this regard is Cook vs. HSBC North America [4], a credit scoring case where the system used as a relevant variable the applicants places of residence, ultimately favoring "white" areas and discriminating against members of ethnic minorities. Those "subtly discriminatory" variable associations (such as zip code/ethnic background) combined with the practical difficulties of accessing relevant information on how an AI system associated different variables, meant that the right to access evidence and courts (as a fair trial safeguard) were under serious threat. To remedy this, regulators across the world and in the EU sought to answer two main questions: 1. which evidence do litigants need to have access to in order to effectively prove causation?; 2. once that evidence has been identified, how should legal procedures be (re)designed to open the victims' access to it?

10.4.1 Using causal tools to establish causal evidence

Regarding the first question, the emerging, but not yet consolidated, global AI liability case law reveals an interesting trend. Although many judicial instances can be cited as examples, for the purpose of this article, we shall highlight three cases that we view as illustrative of the 'new approach' to proving causation in AI-related disputes. These cases are Pickett [5] (dealing with a DNA matching system - TrueAllele - used by police authorities to track down harm-doers),

[3] In the EU, the concept of rule of law is understood to include the following principles: legality, legal certainty, prohibition of arbitrariness of the executive powers, independent and impartial courts, effective judicial review, including respect for fundamental rights and equality before the law. See Communication from the Commission to the European Parliament and the Council "A new EU Framework to strengthen the Rule of Law," COM(2014) 158 final, at 4.

[4] US District Court for the Northern District of Illinois, 21 March 2014, County of Cook v. HSBC North America Holdings Inc et al., 1:2014cv02031.

[5] Superior Court of New Jersey (Appellate Division), 2 February 2021, State of New Jersey v. Corey Pickett, Docket N° A-4207-19T4.

Loomis [6] (dealing with COMPAS, a recidivism-predicting system used by courts) and Ewert [7] (also dealing with the use of recidivism-predicting systems by Canadian correctional services). In all three cases, the plaintiffs argued that the automated decisions were inaccurate because they were unfair that is, contained unfair biases: gender in Pickett and Loomis, ethnic background in Ewert. To uncover the bias-conducive variable association (i.e the causal link), the plaintiffs requested that the systems be reverse engineered. This was hardly possible. For example, in Pickett, independent experts confirmed that reverse engineering would take up to 8,5 years to be completed [8]. In the face of the practical unfeasibility of reverse-engineering, the court in Pickett (and in Loomis) turned to general expertise, as a faute de mieux solution: the lack of direct evidence (reverse-engineering) able to reveal the presence of an unfair bias, was "compensated" by the recourse to already existing expertise assessing a system's functionalities in general. If the majority of experts agreed that a system, like TrueAllele in Pickett or COMPAS in Loomis, was generally well-performing (i.e. was unbiased and therefore accurate), the courts would be inclined to accept that, in the disputes they were called to resolve, it could be presumed that the systems concerned had made unbiased decisions.

Hence, the role of experts is to assess the strength of the causal link between sensitive variables and the decision (A zero causal effect indicates absence of discrimination) in presence of different causal structures (Section 10.2) which can lead to different types of bias. A possible approach would be to identify the causal graph to reveal the causal relations between variables and then use causal notions of fairness (Section 10.3) to assess discrimination. A suggested procedure to identify the causal graph is to first use a causal discovery algorithm (e.g. PC [212]). Then, seek the input of experts in the domain of application to adjust the discovered graph (e.g. adding/removing causal links, enforcing assumptions, etc.). The input of experts can be useful also to clarify the role of each variable, in particular, classifying mediator variables into explaining (leading to justifiable discrimination) and proxy (leading to unjustifiable discrimination) variables. This is essential to select the suitable causal fairness metric (Section 10.3) to use.

10.4.2 But-for test using counterfactuals

From the perspective of procedural fairness and the mediator structure model, this trend is, of course, open to criticism. First, the general opinions of experts on the accuracy of a system are not as probative as direct evidence (reverse engineering) able to provide highly reliable information on the mediator association having led to a discriminatory outcome. Second, the inability to prove causation through reliable evidence seems to have given way to a peculiar application of the so-called but-for test. In principle, this test translates to the deployment of counterfactual reasoning seeking to determine if a harm would have been suffered, had an alleged cause not occurred. In the Cook vs HSBC case (credit scoring) e.g., a standard application of said test would translate to determining if the same loan applicants would have been approved, if the

[6] Supreme Court of Wisconsin, 13 July 2016 (decided), State of Wisconsin v. Eric L. Loomis, 881 N.W. 2d 749 (2016) 2016 WI 68.

[7] Ewert vs. Canada, 2018 SCC 30, File n° 37233, 13 June 2018.

[8] See Superior Court of New Jersey (Appellate Division), 2 February 2021, State of New Jersey v. Corey Pickett, Docket N° A-4207-19T4, at 17.

system had not taken their places of residence as a relevant variable. However, the cases cited in this section (in particular Pickett and Loomis), reveal a slight shift in the application of the but-for test. In "ordinary" disputes (non-AI related) cases, this test seeks to answer a question of factive causal association: would an outcome be the same (or different) without certain facts (address, gender, age, etc) in the causal structure? In AI-related disputes, the but/for test answers a question of (human) reliance on AI output, the relevant (causal) issue being if a human decision based on AI would have been the same or different, had the AI not been used at all. In this case, statistical causality tools can be applied to detect discrimination in data reflecting previous hiring or loan-granting practices in the company concerned. If the association between the sensitive attribute and the outcome is detected, then one can conclude that the decision would be the same without algorithmic assistance. This brings the focus of attention from the AI system (and its architects) to the general practice in the company. Here, again, causality can help to distinguish between spurious association, explainable disparity, or discrimination. On the other hand, if the data with ingrained discrimination is the same as that used to train the algorithm, compliance with AI designing guidelines can be further scrutinized. Finally causality tools provide mathematical expressions to capture the intangible concept of counterfactual [213, 214] very useful to directly check the but-for test.

This allows us to raise the second issue mentioned above: should systems of evidence include a right to access/to request disclosure of evidence?

10.4.3 Disclosing causal evidence to victims of discrimination

From a procedural fairness perspective, this right seems paramount for a victim of algorithmic discrimination to at least have a shot at requesting the 'lifting of the opacity veil' that might cover a causal chain [215]. In the EU, recent regulatory developments seemed - on the surface at least - to move toward the recognition of such a right. First came the AI Act [9] - a horizontal, across-the-board legislation which makes two important contributions. On the one hand, it includes a four-level taxonomy of risks-of-harm related to AI systems: non-high, limited, high and unacceptable. On the other hand, and against the backdrop of said risk-taxonomy, the AI Act includes a set of technical standards (transparency, data governance, risk-mitigation strategy etc) targeting high-risk AI systems, used in mainly eight market sectors [10]. To complement the AI Act and to afford procedures designed for the compensation of harm associated with high-risk systems, the AI Liability Directive (AILD) [11] came next. This instrument establishes a system of evidence which grants victims the right to request disclosure of evidence. By virtue of the AILD, if the defendant (a programmer or user) refused to disclose the evidence requested by the victim or if, upon disclosure, a national or EU court found that the evidence was probative and plausible, the defendant would be presumed responsible for the harm (e.g. discrimination) suffered by the

[9] Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (AI Act) and amending certain Union legislative acts, COM(2021) 206 final.

[10] The 'high-risk' sectors are listed in Annex III of the AI act and include Employment, education, healthcare, transport, energy, public sector (including asylum, migration, border controls, judiciary and social security services), defence and security, finance, banking, and insurance.

[11] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to Artificial Intelligence (AI Liability Directive) COM(2022)496 final.

claimant. It should however be stressed that the evidence a victim can ask disclosure of under the AILD does not include the evidence flagged as 'necessary' (i.e. expertise) in the cases cited earlier. The AILD allows the disclosure of evidence so long as that evidence pertains to the defendant's compliance with the technical standards listed in the AI Act. In other words, the defendant would not be asked to provide information (e.g. access to the code, reverse-engineering, when feasible) able to support a proper causal analysis. They would be asked to - merely - provide information confirming that they complied with, say, their duty for human control and oversight. The reason for this is, no doubt, that the AILD relies on the assumption that if harm (like discrimination) does occur, it is because the AI Act had not been fully observed. In doing so, the AILD narrows down the scope of the evidentiary debate in the sense that the parties in future AI discrimination cases, will not seek to be called to uncover the actual casual structure underlying discriminatory AI output, but to identify the human agent who had failed to meet a legally prescribed duty of care.

10.5 Practical Considerations for Using Causality for Fairness

In the previous sections, we illustrated the situations where the causality approach is relevant for evaluating fairness and how it can be attained using causal fairness notions and approximation techniques. Despite the apparent advantages, the applicability of the causal framework is limited because of its reliance on prior knowledge and often untestable assumptions. Many causal requirements can be achieved by applying a specific experiment design (ideally, random assignment). However, in fairness scenarios, it is often not a plausible option. Therefore, discrimination is usually evaluated from observational data. Here, we will list some requirements for applying causal inference that are most relevant for fairness applications.

10.5.1 Possibility for Intervention

In fairness estimation, the sensitive attribute is considered to be the exposure or treatment attribute. The goal is to measure its impact on the outcome. Most definitions of causal effect are based on a notion of intervention or manipulation of a cause variable (exposure) [216, 217]. This makes it hard to justify causal claims related to non-manipulable quantities, such as sensitive attributes, for example, race or gender. Some approaches in the literature suggest shifting attention from an actual manipulation to changes in perception [95]. For example, instead of changing the gender of candidates to estimate the effect on a hiring decision, the researcher could manipulate the perception of gender by the employer. It could be easily done by submitting two analogous resumes but varying the name or title of the candidate. This approach corresponds to the methodology applied in social experiments on the impact of race or gender of an applicant on hiring decisions [218]. [216] further differentiate immutable sensitive attributes into those that are randomized at birth (biological sex) and those that are not (race, social gender). This distinction is important when estimating the causal effect of the sensitive attribute. If the sensitive attribute is randomized, then its causal effect on an outcome can be estimated just by comparing exposure levels. For example, it is possible to estimate the total causal effect of biological sex

by taking the observed differences in the outcome between men and women [216]. In contrast, race is not random, but depends on many ancestral factors. For this reason, even at the biological level, estimating the effect of race is more complicated and requires information about the causal structure of the covariates. However, these types of estimation are relevant in medical scenarios, where the independence between the sensitive attribute and the outcome (for example, the probability of a disease) cannot be reasonably assumed. In the possible discrimination scenarios [216], similarly, to [95] shift attention to the direct effect of the *perceived* gender or race on the decision.

10.5.2 Causal assumptions

The SUTVA [112] (Stable Unit Treatment Value Assumption) entails the requirements of no interference and consistency. No interference assumption requires that the interaction between individuals does not influence the effect of the sensitive attribute on the outcome. The likelihood of interaction and feedback loops is high in social sciences research in general and calls for a clear discussion and restricted interpretations of causal estimation [50]. Fairness is usually measured in a social context. Therefore, the possibility of interaction should be carefully evaluated. Using the hiring example, the violation of the SUTVA requirement would occur in a situation where hiring more participants of one political spectrum increases the likelihood of privileging the same political spectrum in future hiring decisions. Such a scenario is plausible because current employees may favor those who have political beliefs similar to their own. The assumption of consistency requires that each treatment level leads to the same potential outcomes [219]. In fairness evaluation, treatment is replaced by the sensitive attribute, which is often a social construct such as race or gender. Identifying the causal effect of gender on hiring can be problematic if gender itself does not have a consistent effect on hiring. For example, only women with a certain level of "femininity" are discriminated against. This scenario cannot be excluded and should be considered if a fine-grained causal analysis is a goal of a study. In summary, SUTVA assumptions are likely to be violated in fairness scenarios, however, causal approaches can still be applied if the results are interpreted with caution. Some methods to identify the causal effect under the violations of SUTVA are discussed here [220].

Ignorability [112] assumption requires that the sensitive attribute and the outcome are independent given the observable variables. In other words, no unobserved variables create a significant link between the sensitive attribute and the outcome. In fairness evaluation, the presence of such a link could mean that the portion of discrimination is, in fact, a spurious effect induced by the confounder. For example, if the education confounder is not present in the data, the confounding effect cannot be controlled. As a result, it is not possible to estimate the causal effect of political belief on the hiring decision that is separate from the effect of education. Unobserved confounders are not likely for immutable sensitive attributes such as sex or race. These sensitive attributes are unlikely to have a temporally prior cause. However, the noise terms can still be not independent between the sensitive attribute and the outcome. [205] point out, the implications of assuming ignorability, when using causal counterfactuals. Following the reasoning by [205], in the case of college admission (Figure **??**), an average male who applied

to the technical profession could be counterfactually exchanged with an average woman who applied to the same profession. However, given the social expectations tied to gender roles, a woman applying to a technical profession is likely to be more motivated and hard-working than an average male with the same professional goals.

Positivity [112] is violated if some of the combinations of a sensitive attribute and a covariate have zero probability. Violations of positivity can be deterministic or random [221]. For example, positivity would be violated if a certain level of education always corresponds to liberal political beliefs. In this scenario, the positivity violation would most likely be random. It is unlikely that certain education would have a deterministic relationship on political beliefs. In the random case, statistical methods are available for analysis under violation of positivity [221]. However, consider a case where having a Harvard degree is considered an explanatory mediator between ethnicity and hiring. Certain ethnicities may have zero probability of having obtained a Harvard degree due to long-term discrimination and poverty. In this case, it should be reconsidered if a specific Harvard degree is essential for the job considered despite the potential exclusion and disparate impact.

The identifiability of path-specific effects in the presence of multiple mediators requires the absence of causal links between the mediators [222]. Evaluating path-specific effects is particularly important to understand the mechanism of the effect of the sensitive attribute on the outcome. As outlined earlier (Section 10.3), the effect can be deemed justifiable or discriminatory depending on the mediating variables on the path. However, the link between two or more mediators is likely in fairness scenarios. For example, consider the case where race and hiring decisions are mediated by social status (redlining) and education (explaining variable). It is very likely that the level of education is influenced by social status. In this case, the indirect effect through social status and education separately is not identifiable. Work by [222] proposed a method based on the treatment of multiple mediators together. In some cases, this method can help identify individual indirect effects in the presence of causal links between mediators.

10.5.3 Availability of Causal Graph

One of the most significant restrictions for using causality is knowledge of the relationship between variables in the form of a directed acyclic graph (DAG) [12] [46]. The research by [223] shows a significant disagreement between estimations of causal fairness notions due to slight differences in the causal structure. The availability of DAG is particularly important in the presence of collider structures, because including a collider in a conditioning set induces bias in measuring causal effect [56]. DAG is also important for the evaluation of path-specific effects, important for distinguishing redlining and explaining variables in fairness scenarios.

The causal structure (or causal graph) can be obtained by consulting domain experts or learning from observational data. Both approaches have their own limitations. Domain experts can disagree or have biased assumptions. Learning from the data often requires additional

[12] The DAG is subject to further assumptions of causal Markov condition, causal faithfulness, and causal sufficiency. Causal Markov condition, causal faithfulness, and Causal sufficiency together encode the same requirements as SUTVA and Ignorability in the potential outcome framework, therefore, will not be discussed separately.

assumptions on the distribution of the data, functional relationships, the relations of exogenous unobserved variables, and the informed choice of the learning algorithm. The research by [223] shows how different algorithms to recover the causal structure yield different results when applied to the same data set. Learning causal relationships from observational data alone may not be realistic [224]. However, the combination of causal discovery and expert knowledge could give more reliable results.

11

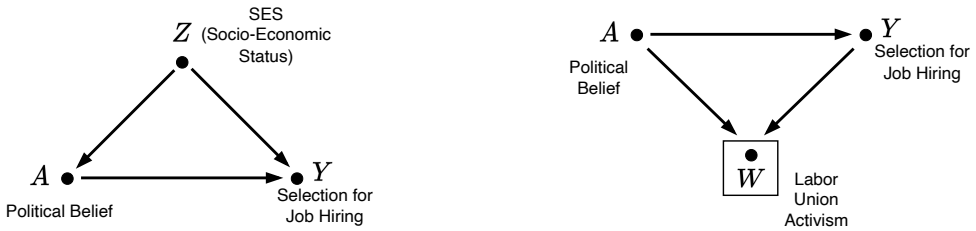
Dissecting Causal Biases

11.1 Introduction

In this chapter, we focus on a class of biases, which we call causal biases, that arise from the way data is generated and/or collected. We make a distinction between discrimination and bias. We use the term discrimination to refer to *the unjust or prejudicial treatment of different categories of people, on the basis of race, age, gender, disability, religion, political belief, etc.*. Whereas the term bias is used to refer to *the deviation of the expected value from the quantity it estimates.* We use tools from the field of causality [46, 225] to characterize causal biases and disentangle them from discrimination.

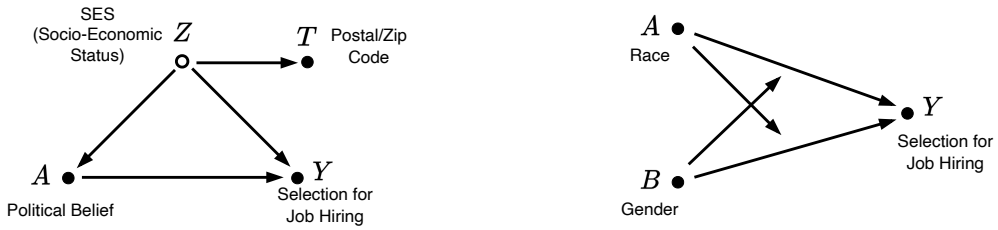
11.1.1 Types of bias

Measurement of discrimination without taking into consideration the causal structure underlying the relationships between variables may lead to misleading conclusions. That is, a biased estimation of discrimination. In extreme cases, such as Simpson's paradox, the bias may lead to reversing the conclusions (e.g. the biased estimation indicates a positive discrimination, while the unbiased estimation is actually a negative discrimination). We will be considering the types of bias based on Collider and Confounder causal structures described in section 10.2. We also add a measurement bias structure (Paragraph 11.1.2, Figure 11.2(a)) and interaction structure(Paragraph 11.1.3,Figure 11.2(b)).



((a)) Confounding bias example. ((b)) Collider bias example.

Figure 11.1: Confounding and colliding bias.



((a)) Measurement bias example. ((b)) Interaction Bias example.

Figure 11.2: Measurement and interaction bias.

11.1.2 Measurement Bias

The third type of bias, measurement bias, is due to the use of a proxy variable to estimate discrimination instead of an ideal but unmeasurable variable. Consider a third variant of the same job selection example having the causal graph of Figure 11.2(a). Unlike in the causal graph of confounding bias (Figure 11.1(a)), the confounder variable Z is unmeasurable (empty bullet instead of a filled one). In practice, it is difficult to find a variable that represents accurately the socio-economic status (salary, possessions, etc.). Being unmeasurable, Z cannot be used to estimate discrimination while blocking the confounding path through Z. For practical reasons, the (measurable) variable T representing the postal/zip code of the candidate's address can be used instead. T is considered a proxy of Z as it is highly correlated with (but not identical to) Z[1]. Using variable T as a proxy to measure Z may lead to an additional bias, we call measurement bias.

11.1.3 Interaction Bias

Interaction bias is observed when two causes of the outcome interact with each other, making the joint effect smaller or greater than the sum of individual effects. Consider the same job hiring example but where two sensitive attributes, political belief (liberals vs. conservatives), and gender have an effect on the hiring decision. In the presence of interaction between political belief and gender, statistical disparity will not accurately measure the individual effects of Political Belief and Gender even if no confounding condition is satisfied. For example, it is possible to observe a situation where statistical parity is almost satisfied for both individual sensitive variables, but the intersectional sensitive group is discriminated [6]. Following our previous example, we

[1]The candidate's address gives a strong indicator of the socio-economic status.

would define liberal females as an unprivileged intersectional group and conservative males as a privileged intersectional group. In the presence of interaction, the discrimination against liberal females is not equal to the sum of discrimination against conservative and females individually. Additionally, the average discrimination value for liberals or females, as measured by statistical disparity, will also be biased, as it does not take into account the interaction between the two sensitive variables.

11.1.4 Notation and preliminaries

Variables are denoted by capital letters. In particular, A is used for the sensitive variable (e.g., gender, race, age) and Y is used for the outcome of the automated decision system (e.g., hiring, admission, releasing on parole). Small letters denote specific values of variables (e.g., $A = a'$, $W = w$). Bold capital and small letters denote sets of variables and sets of values, respectively.

The Back door formula and Average Causal Effect (ACE)

Assuming Z is the only confounder of A and Y, the back door formula can be used to control for the confounding effect the Back door formula is expressed as:

DEFINITION 11.1.1.

$$\mathbb{P}(Y|do(A = a)) = \sum_{z \in Z} \mathbb{P}(Y|A = a, Z = z)\mathbb{P}(Z = z) \tag{11.1}$$

Average causal affect (ACE) is defined as:

DEFINITION 11.1.2.

$$ACE(Y, A) = \mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)). \tag{11.2}$$

Statistical Disparity

Statistical disparity (Equation 3.1) is a biased estimation of the discrimination in the presence of a confounder variable, Z, between A and Y as it does not filter out the spurious effect due to the confounding. For the sake of the proofs, we define the following variant of statistical disparity:

DEFINITION 11.1.3.

$$StatDisp(Y, A)_Z = \sum_{z \in Z} \big(\mathbb{P}(y_1|a_1, z) - \mathbb{P}(y_1|a_0, z)\big).\mathbb{P}(z). \tag{11.3}$$

Notice that if Z d-separates[2] A and Y, $StatDisp(Y, A)_Z$ coincides with the average causal effect ACE (Equation 11.2).

[2]For the definition of d-separation, we refer the reader to Definition 1.2.3 in [46].

11.1.5 Previous results used in the proofs

Consider a pair of variables X and Y. The variance of a variable X, ${\sigma_x}^2$, is a measure of dispersion which quantifies how far a set of values deviate from their mean and is defined as: ${\sigma_x}^2 = \mathbb{E}[X - \mathbb{E}[X]]^2$. Covariance of X and Y, σ_{xy}, is a measure of the joint variability of two random variables and is defined as: $\sigma_{xy} = \mathbb{E}[[X - \mathbb{E}[X][Y - \mathbb{E}[Y]]]$. Assuming a linear relationship between X and Y (X is the predictor variable, while Y is the response variable), the regression coefficient of Y given X, β_{yx}, represents the slope of the regression line in the prediction of Y given X ($\frac{\partial}{\partial x}\mathbb{E}[Y|X = x]$) and is equal to $\beta_{yx} = \frac{\sigma_{xy}}{{\sigma_x}^2}$. Correlation coefficient ρ_{yx}, however, represents the slope of the least square error line in the prediction of Y given X. The relationships between σ_{yx}, β_{yx}, and ρ_{yx} are as follows:

$$\beta_{yx} = \frac{\sigma_{yx}}{\sigma_x^2} = \rho_{yx}\frac{\sigma_y}{\sigma_x}$$
$$\rho_{yx} = \rho_{xy} = \frac{\sigma_{yx}}{\sigma_x\sigma_y} = \beta_{yx}\frac{\sigma_x}{\sigma_y} = \beta_{xy}\frac{\sigma_y}{\sigma_x}$$

Partial regression coefficient, $\beta_{yx.z}$, represents the slope of the regression line of Y on X when we hold variable Z constant ($\frac{\partial}{\partial x}\mathbb{E}[Y|X = x, Z = z]$). A well known result by Cramer [226] allows to express $\beta_{yx.z}$ in terms of covariance between pairs of variables [227]:

$$\beta_{yx.z} = \frac{{\sigma_z}^2\sigma_{xy} - \sigma_{yz}\sigma_{zx}}{{\sigma_x}^2{\sigma_z}^2 - {\sigma_{xz}}^2} \tag{11.4}$$

For standardized variables (all variables are normalized to have a zero mean and a unit variance), the partial regression coefficient has a simpler expression since $\beta_{yx} = \sigma_{yx}$:

$$\beta_{yx.z} = \frac{\sigma_{xy} - \sigma_{yz}\sigma_{zx}}{1 - {\sigma_{xz}}^2} \tag{11.5}$$

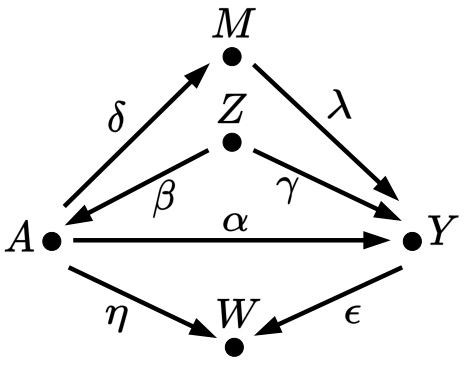


Figure 11.3: Causal graph with linearly related variables. Arrow labels represent linear regression coefficients.

Another known result by Wright and Pearl [60, 227] allows to represent the covariance of two variables in terms of the regression coefficients of the different paths (causal and non-causal, but not passing through any collider variable) between those two variables. More precisely, σ_{yx} is equal to the sum of the regression coefficients of every path between x and y, weighted by the variance of the root variable of each path. For instance, in Figure 11.3, $\sigma_{ya} = {\sigma_a}^2\alpha + {\sigma_z}^2\beta\gamma + {\sigma_a}^2\delta\lambda$.

Notice that the coefficients η and ϵ are not included as the path $A \rightarrow W \leftarrow Y$ is not $d-$connected (W is a collider variable). For standardized variables, the expression is simpler as all variables are normalized to have a unit variance. For the same example (Figure 11.3), $\sigma_{ya} = \alpha + \beta\gamma + \delta\lambda$. For linear models, regression coefficients can be interpreted causally. For instance, using the same example of Figure 11.3, α represents the direct causal effect of A on Y.

11.2 Confounding bias

Confounding bias occurs when both the sensitive variable and the outcome have a common cause, the counfounder variable (Figure 11.4). Consequently, the mechanism of selecting samples from the two groups (protected and privileged) is not independent of the outcome. This creates a bias when measuring the causal effect of the sensitive attribute on the outcome.

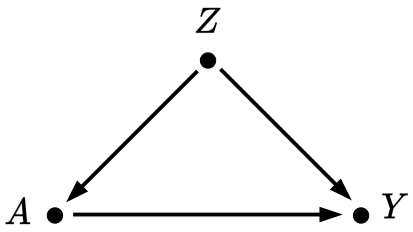


Figure 11.4: Simple confounding structure

11.2.1 Binary Model Case

For a concise notation, let y_1 and y_0 denote the propositions $Y = 1$ and $Y = 0$, respectively, and the same for the variables A and Z. For instance, $\mathbb{P}(Y = 1|A = 0)$ is written simply as $\mathbb{P}(y_1|a_0)$.

Statistical disparity [228] between groups $A = 0$ and $A = 1$, denoted as $StatDisp(Y, A)$, is the difference between the conditional probabilities: $\mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0)$. In presence of a confounder variable, Z, between A and Y, statistical disparity is a biased estimation of the discrimination as it does not filter out the spurious effect due to the confounding.

DEFINITION 11.2.1. Confounding bias is defined as[3]:

$$ConfBias(Y, A) = StatDisp(Y, A) - ACE(Y, A) \tag{11.6}$$

where $ACE(Y, A)$ is the causal effect of A on Y (Definition 11.2) and $StatDisp(Y, A)$ is a statistical relationship between A and Y (Definition 3.1). For the simple confounding structure of Figure 11.4, ACE coincides with $StatDisp_Z(Y, A)$ (Definition 11.1.3).

THEOREM 11.1. The difference in discrimination due to confounding bias is equal

[3]In this chapter, bias is defined by substracting the correct value of discrimination from the biased estimation.

to:

$$\begin{aligned} \mathit{ConfBias}(Y,A) = (1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1)) \\ \times (\alpha - \beta - \gamma + \delta + \frac{\gamma}{\mathbb{P}(a_1)} - \frac{\delta}{\mathbb{P}(a_1)}) \end{aligned} \tag{11.7}$$

where α, β, γ, and δ denote, respectively, $\mathbb{P}(y_1|a_0, z_0), \mathbb{P}(y_1|a_0, z_1), \mathbb{P}(y_1|a_1, z_0)$, and $\mathbb{P}(y_1|a_1, z_1)$.

Proof. Let $\mathbb{P}(z_1) = \epsilon$ ($\epsilon \in]0,1[$) and hence $\mathbb{P}(z_0) = 1 - \epsilon$. Similarly, let $\mathbb{P}(a_1) = \lambda$ and hence $\mathbb{P}(a_0) = 1 - \lambda$. Let $\mathbb{P}(y_1|a_0, z_0) = \alpha$, $\mathbb{P}(y_1|a_0, z_1) = \beta$, $\mathbb{P}(y_1|a_1, z_0) = \gamma$, and $\mathbb{P}(y_1|a_1, z_1) = \delta$. Finally, let $\mathbb{P}(z_0|a_0) = \tau$. The remaining conditional probabilities of Z given A are equal to the following:

$$\mathbb{P}(z_1|a_0) = 1 - \mathbb{P}(z_0|a_0) = 1 - \tau \tag{11.8}$$

$$\begin{aligned} \mathbb{P}(z_1|a_1) &= \frac{\mathbb{P}(z_1) - \mathbb{P}(z_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\ &= \frac{\epsilon - (1-\tau)(1-\lambda)}{\lambda} \\ &= \frac{\epsilon - 1 + \tau + \lambda - \tau\lambda)}{\lambda} \end{aligned} \tag{11.9}$$

$$\begin{aligned} \mathbb{P}(z_0|a_1) &= \frac{\mathbb{P}(z_0) - \mathbb{P}(z_0|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\ &= \frac{(1-\epsilon) - \tau(1-\lambda)}{\lambda} \\ &= \frac{1 - \epsilon - \tau + \tau\lambda}{\lambda} \end{aligned} \tag{11.10}$$

Equation (11.8) follows from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(a_i|X) = 1$. Equations (11.9) and (11.10) follow from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(X|u_i)\mathbb{P}(u_i) = \mathbb{P}(X)$. $\mathit{StatDisp}(Y,A)$ can then be expressed in terms of the above parameters:

$$\begin{aligned} \mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) &= \sum_{z \in Z} (\mathbb{P}(y_1|a_1, z)\mathbb{P}(z|a_1) - \mathbb{P}(y_1|a_0, z)\mathbb{P}(z|a_0) \\ &= \mathbb{P}(y_1|a_1, z_0)\mathbb{P}(z_0|a_1) - \mathbb{P}(y_1|a_0, z_0)\mathbb{P}(z_0|a_0) \\ &\quad + \mathbb{P}(y_1|a_1, z_1)\mathbb{P}(z_1|a_1) - \mathbb{P}(y_1|a_0, z_1)\mathbb{P}(z_1|a_0) \\ &= \gamma\left(\frac{1 - \epsilon - \tau\lambda}{1 - \lambda}\right) - \alpha\tau + \delta\left(\frac{\epsilon - \lambda + \tau\lambda}{1 - \lambda}\right) - \beta(1 - \tau) \end{aligned}$$

$ACE(Y, A)$, on the other hand can be expressed as follows:

$$\begin{aligned}\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)) &= \sum_{z\in Z}(\mathbb{P}(y_1|a_1, z) - \mathbb{P}(y_1|a_0, z))\mathbb{P}(z)\\ &= \mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_0, z_0))\mathbb{P}(z_0)\\ &\quad + \mathbb{P}(y_1|a_1, z_1) - \mathbb{P}(y_1|a_0, z_1))\mathbb{P}(z_1)\\ &= (\gamma - \alpha)(1 - \epsilon) + (\delta - \beta)\epsilon\end{aligned}$$

Confounding bias is then equal to:

$$\begin{aligned}StatDisp(Y, A) - ACE(Y, A) &= \mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) - (\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)))\\ &= \gamma\big(\frac{1 - \epsilon - \tau\lambda}{1 - \lambda}\big) - \alpha\tau + \delta\big(\frac{\epsilon - \lambda + \tau\lambda}{1 - \lambda}\big) - \beta(1 - \tau)\\ &\quad - \big((\gamma - \alpha)(1 - \epsilon) + (\delta - \beta)\epsilon\big)\\ &= \frac{\gamma}{\lambda} - \frac{\gamma\epsilon}{\lambda} - \frac{\gamma\tau}{\lambda} + \gamma\tau - \alpha\tau + \frac{\delta\epsilon}{\lambda} - \frac{\delta}{\lambda} + \frac{\delta\tau}{\lambda}\\ &\quad + \delta - \delta\tau - \beta + \beta\tau - \gamma + \alpha + \epsilon\gamma - \alpha\epsilon - \delta\epsilon + \beta\epsilon\\ &= \big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big) - \tau\big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big)\\ &\quad - \epsilon\big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big)\\ &= \big(1 - \tau - \epsilon\big)\big(\alpha + \delta - \beta - \gamma + \frac{\gamma}{\lambda} - \frac{\delta}{\lambda}\big)\end{aligned}$$

□

For the specific case of equal proportions between sensitive groups (e.g. no under or over representation of a certain sensitive group), confounding bias can be characterized by a simpler closed-form expression.

THEOREM 11.2. Assuming that $\mathbb{P}(a_0) = \mathbb{P}(a_1) = \frac{1}{2}$, the difference in discrimination due to confounding bias is equal to:

$$ConfBias(Y, A) = (1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1))(\alpha - \beta + \gamma - \delta) \tag{11.11}$$

where α, β, γ, and δ are defined similarly to Theorem 11.1.

Proof. Let $\mathbb{P}(z_1) = \epsilon$ ($\epsilon \in]0, 1[$) and hence $\mathbb{P}(z_0) = 1 - \epsilon$. And let $\mathbb{P}(y_1|a_0, z_0) = \alpha$, $\mathbb{P}(y_1|a_0, z_1) = \beta$, $\mathbb{P}(y_1|a_1, z_0) = \gamma$, and $\mathbb{P}(y_1|a_1, z_1) = \delta$. Finally, let $\mathbb{P}(z_0|a_0) = \tau$. The remaining conditional

probabilities of Z given A are equal to the following:

$$\mathbb{P}(z_1|a_0) = 1 - \mathbb{P}(z_0|a_0) = 1 - \tau \tag{11.12}$$

$$\mathbb{P}(z_1|a_1) = \frac{\mathbb{P}(z_1) - \mathbb{P}(z_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)}$$

$$= 2\epsilon + \tau - 1 \tag{11.13}$$

$$\mathbb{P}(z_0|a_1) = 1 - \mathbb{P}(z_1|a_1) \tag{11.14}$$

$$= 2 - 2\epsilon - \tau$$

Equations (11.12) and (11.14) follow from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(a_i|X) = 1$. Equation (11.13) follows from the fact that, given u_i events are exhaustive and mutually exclusive, $\sum_i \mathbb{P}(X|u_i)\mathbb{P}(u_i) = \mathbb{P}(X)$. $StatDisp(Y, A)$ can then be expressed in terms of the above parameters:

$$\begin{aligned}
\mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) &= \sum_{z\in Z}(\mathbb{P}(y_1|a_1, z)\mathbb{P}(z|a_1) - \mathbb{P}(y_1|a_0, z)\mathbb{P}(z|a_0) \\
&= \mathbb{P}(y_1|a_1, z_0)\mathbb{P}(z_0|a_1) - \mathbb{P}(y_1|a_0, z_0)\mathbb{P}(z_0|a_0) \\
&\quad + \mathbb{P}(y_1|a_1, z_1)\mathbb{P}(z_1|a_1) - \mathbb{P}(y_1|a_0, z_1)\mathbb{P}(z_1|a_0) \\
&= \gamma(2 - 2\epsilon - \tau) - \alpha\tau + \delta(2\epsilon + \tau - 1) - \beta(1 - \tau) \\
&= \tau(-\alpha + \beta - \gamma + \delta) + 2\epsilon(\delta - \gamma) + 2\gamma - \delta - \beta
\end{aligned}$$

$ACE(Y, A)$, on the other hand can be expressed as follows:

$$\begin{aligned}
\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)) &= \sum_{z\in Z}(\mathbb{P}(y_1|a_1, z) - \mathbb{P}(y_1|a_0, z))\mathbb{P}(z) \\
&= \mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_0, z_0))\mathbb{P}(z_0) \\
&\quad + \mathbb{P}(y_1|a_1, z_1) - \mathbb{P}(y_1|a_0, z_1))\mathbb{P}(z_1) \\
&= (\gamma - \alpha)(1 - \epsilon) + (\delta - \beta)\epsilon
\end{aligned}$$

Confounding bias is then equal to:

$$\begin{aligned}
StatDisp(Y, A) - ACE(Y, A) &= \mathbb{P}(y_1|a_1) - \mathbb{P}(y_1|a_0) - (\mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0))) \\
&= \tau(-\alpha + \beta - \gamma + \delta) + 2\epsilon(\delta - \gamma) + 2\gamma - \delta - \beta \\
&\quad - ((\gamma - \alpha)(1 - \epsilon) + (\delta - \beta)\epsilon) \\
&= \tau(-\alpha + \beta - \gamma + \delta) + 2\epsilon\delta - 2\epsilon\gamma + 2\gamma - \delta - \beta \\
&\quad - \gamma + \gamma\epsilon + \alpha - \alpha\epsilon - \delta\epsilon + \beta\epsilon \\
&= \tau(-\alpha + \beta - \gamma + \delta) + \epsilon(2\delta - 2\gamma + \gamma - \alpha - \delta + \beta) \\
&\quad + 2\gamma - \delta - \beta - \gamma + \alpha \\
&= \tau(-\alpha + \beta - \gamma + \delta) + \epsilon(-\alpha + \beta - \gamma + \delta) + \alpha - \beta + \gamma - \delta \\
&= (1 - \tau - \epsilon)(\alpha - \beta + \gamma - \delta)
\end{aligned}$$

□

11.2.2 Linear Model Case

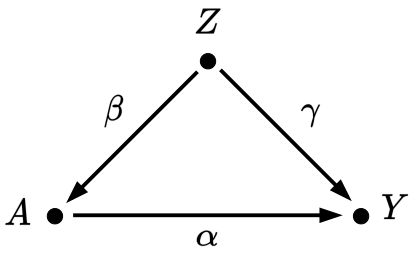


Figure 11.5: Confounding structure in linear model

THEOREM 11.3. Let A, Y, and Z variables with linear regressions coefficients as in Figure 11.5 which represents the basic confounding structure. The confounding bias can be expressed in terms of covariances of pairs of variables as follows:

$$ConfBias(Y, A) = \frac{\sigma_{za}\sigma_{yz} - \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{za}^2}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \tag{11.15}$$

Confounding bias can also be expressed in terms of the linear regression coefficients as follows:

$$ConfBias(Y, A) = \frac{{\sigma_z}^2}{{\sigma_a}^2}\beta\gamma \tag{11.16}$$

Proof. For Equation (11.15),

$$\begin{aligned}
ConfBias(Y, A) &= \beta_{ya} - \beta_{ya.z} \\
&= \frac{\sigma_{ya}}{\sigma_a^2} - \frac{\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\frac{\sigma_{ya}}{\sigma_a^2}(\sigma_a^2\sigma_z^2 - \sigma_{za}^2) - (\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za})}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\cancel{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_a^2}\sigma_z^2 - \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{za}^2 - \cancel{\sigma_z^2\sigma_{ya}} + \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\sigma_{za}\sigma_{yz} - \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{za}^2}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2}
\end{aligned}$$

For Equation (11.16),

$$\begin{aligned}
ConfBias(Y,A) &= \beta_{ya} - \beta_{ya.z} \\
&= \frac{\sigma_{ya}}{\sigma_a^2} - \frac{\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\sigma_a^2\alpha + \sigma_z^2\beta\gamma}{\sigma_a^2} - \frac{\sigma_z^2(\sigma_a^2\alpha + \sigma_z^2\beta\gamma) - (\sigma_z^2\gamma + \sigma_z^2\beta\alpha)(\sigma_z^2\beta)}{\sigma_a^2\sigma_z^2 - (\sigma_z^2\beta)^2} \\
&= \frac{\cancel{\sigma_a^2}\alpha}{\cancel{\sigma_a^2}} + \frac{\sigma_z^2\beta\gamma}{\sigma_a^2} - \frac{\sigma_z^2\sigma_a^2\alpha + \cancel{\sigma_z^4\beta\gamma} - \cancel{\sigma_z^4\beta\gamma} - \sigma_z^4\beta^2\alpha}{\sigma_a^2\sigma_z^2 - \sigma_z^4\beta^2} \\
&= \cancel{\alpha} + \frac{\sigma_z^2\beta\gamma}{\sigma_a^2} - \frac{\cancel{\alpha(\sigma_z^2\sigma_a^2 - \sigma_z^4\beta^2)}}{\cancel{\sigma_a^2\sigma_z^2 - \sigma_z^4\beta^2}} \\
&= \frac{\sigma_z^2}{\sigma_a^2}\beta\gamma
\end{aligned}$$

□

COROLLARY 11.4. For standardized variables A, Y, and Z, confounding bias can be expressed in terms of covariances as:

$$ConfBias(Y,A) = \frac{\sigma_{za}\sigma_{yz} - \sigma_{ya}\sigma_{za}^2}{1 - \sigma_{za}^2} \tag{11.17}$$

And in terms of regression coefficient, simply as ([227]):

$$ConfBias(Y,A) = \beta\gamma \tag{11.18}$$

Equations (11.17) and (11.18) can be obtained from Equations (11.15) and (11.16) as $\sigma_z = \sigma_a = 1$.

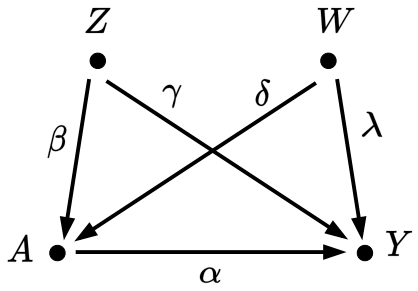


Figure 11.6: Confounding structure with two confounders

THEOREM 11.5. Let A, Y, Z, W variables as in Figure 11.6. Assuming that all variables are standardized and that W and Z are independent, the regression coefficent of Y on A conditioning on Z and W, the confounding bias is equal:

$$ConfBias(Y,A) = \frac{\sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw} - \sigma_{ya}(\sigma_{za}^2 + \sigma_{wa}^2)}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \tag{11.19}$$

And in terms of the regression coefficients:

$$ConfBias(Y, A) = \beta\gamma + \delta\lambda \tag{11.20}$$

Proof. The proof is based on proving that:

$$\beta_{ya.zw} = \frac{\sigma_{ya} - \sigma_{za}\sigma_{yz} - \sigma_{wy}\sigma_{wa}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \tag{11.21}$$

From Cramér [226] (Page 307), we know that the partial regression coefficient can be expressed as:

$$\beta_{ya.zw} = \rho_{ya.zw}\frac{\sigma_{y.zw}}{\sigma_{a.zw}} \tag{11.22}$$

Where $\rho_{ya.zw}$ denotes the partial correlation and $\sigma_{a.zw}, \sigma_{y.zw}$ denote the residual variances. Based on the correlation matrix:

$$\begin{bmatrix} 1 & \rho_{ya} & \rho_{yz} & \rho_{yw} \\ \rho_{ay} & 1 & \rho_{az} & \rho_{aw} \\ \rho_{zy} & \rho_{za} & 1 & \rho_{zw} \\ \rho_{wy} & \rho_{wa} & \rho_{wz} & 1 \end{bmatrix}$$

the partial correlation $\rho_{ya.zw}$ can be expressed in terms of cofactors as follows[4]:

$$\rho_{ya.zw} = -\frac{C_{ya}}{\sqrt{C_{yy}C_{aa}}} \tag{11.23}$$

where C_{ij} denotes the cofactor of the element ρ_{ij} in the determinant of the correlation matrix and are equal to the following:

$$\begin{aligned} C_{ya} = -(&\rho_{ya} - \rho_{ya}\rho_{zw}^2 - \rho_{za}\rho_{yz} - \rho_{wa}\rho_{yw} \\ &+ \rho_{za}\rho_{yw}\rho_{wz} + \rho_{wa}\rho_{yz}\rho_{zw}) \end{aligned} \tag{11.24}$$

$$C_{yy} = 1 - \rho_{zw}^2 - \rho_{za}^2 - \rho_{wa}^2 + 2\rho_{za}\rho_{aw}\rho_{wz} \tag{11.25}$$

$$C_{aa} = 1 - \rho_{zw}^2 - \rho_{zy}^2 - \rho_{wy}^2 + 2\rho_{yz}\rho_{yw}\rho_{wz} \tag{11.26}$$

Residual variances in Equation 11.22 can be expressed in terms of total and partial correlation coefficients as follows [226](Equation 23.4.5 in page 307):

$$\sigma_{y.zw}^2 = \sigma_y^2(1 - \rho_{yz}^2)(1 - \rho_{yw.z}^2)(1 - \rho_{ya.zw}^2) \tag{11.27}$$

$$\sigma_{a.zw}^2 = \sigma_a^2(1 - \rho_{az}^2)(1 - \rho_{aw.z}^2)(1 - \rho_{ay.zw}^2) \tag{11.28}$$

[4]The proof is sketched in https://en.wikipedia.org/wiki/Partial_correlation.

As the last term is the same, we have:

$$\frac{\sigma_{y.zw}}{\sigma_{a.zw}} = \frac{\sigma_y\sqrt{(1-\rho_{yz}^2)(1-\rho_{yw.z}^2)}}{\sigma_a\sqrt{(1-\rho_{az}^2)(1-\rho_{aw.z}^2)}} \tag{11.29}$$

The partial correlation coefficients in Equation 11.29 can be expressed in terms of total correlation coefficients as follows [226] (Equation 23.4.3 in page 306):

$$\rho_{yw.z} = \frac{\rho_{yw} - \rho_{yz}\rho_{wz}}{\sqrt{(1-\rho_{yz}^2)(1-\rho_{wz}^2)}} \tag{11.30}$$

After simple algebraic steps, we obtain:

$$\frac{\sigma_{y.zw}}{\sigma_{a.zw}} = \frac{\sigma_y}{\sigma_a}\frac{\sqrt{1-\rho_{zw}^2-\rho_{yz}^2-\rho_{yw}^2+2\rho_{zy}\rho_{yw}\rho_{wz}}}{\sqrt{1-\rho_{zw}^2-\rho_{az}^2-\rho_{aw}^2+2\rho_{za}\rho_{aw}\rho_{wz}}} \tag{11.31}$$

Finally, $\beta_{ya.zw}$ in Equation 11.22 can be expressed in terms of total correlation coefficients as follows:

$$\beta_{ya.zw} = \frac{\sigma_y}{\sigma_a}\frac{Q}{1-\rho_{zw}^2-\rho_{za}^2-\rho_{wa}^2+2\rho_{za}\rho_{aw}\rho_{wz}} \tag{11.32}$$

where

$$\begin{aligned} Q = {} & \rho_{ya} - \rho_{ya}\rho_{zw}^2 - \rho_{za}\rho_{yz} - \rho_{wy}\rho_{wa} \\ & + \rho_{za}\rho_{yw}\rho_{zw} + \rho_{wa}\rho_{yz}\rho_{zw} \end{aligned}$$

Recall that $\rho_{ya} = \frac{\sigma_{ya}}{\sigma_y\sigma a}$. The formula becomes:

$$\beta_{ya.zw} = \frac{Q}{R} \tag{11.33}$$

Where

$$\begin{aligned} Q = {} & \sigma_{ya}(\sigma_z^2\sigma_w^2 - \sigma_{zw}^2) + \sigma_{yz}(\sigma_{wa}\sigma_{zw} - \sigma_{za}\sigma_w^2) \\ & + \sigma_{wy}(\sigma_{za}\sigma_{zw} - \sigma_{wa}\sigma_z^2) \end{aligned}$$

And

$$\begin{aligned} R = {} & \sigma_a^2\sigma_z^2\sigma_w^2 - \sigma_a^2\sigma_{zw}^2 - \sigma_a\sigma_z\sigma_w^2\sigma_{za}^2 \\ & - \sigma_z^2\sigma_{aw}^2 + 2\sigma_a\sigma_z\sigma_{az}\sigma_{aw}\sigma_{zw} \end{aligned}$$

For standardized variables, $\forall v, \sigma_v = 1$, and hence $\forall u, v\sigma_{uv} = \rho_{uv}$. Equation 11.32 becomes:

$$\beta_{ya.zw} = \frac{Q}{1-\sigma_{zw}^2-\sigma_{za}^2-\sigma_{wa}^2+2\sigma_{za}\sigma_{aw}\sigma_{wz}} \tag{11.34}$$

Where

$$Q = \sigma_{ya}(1 - \sigma_{zw}^2) + \sigma_{yz}(\sigma_{wa}\sigma_{zw} - \sigma_{za}) + \sigma_{yw}(\sigma_{za}\sigma_{zw} - \sigma_{wa})$$

If we further assume that confounders are uncorrelated, that is, $\sigma_{zw} = 0$, then we have the simpler expression:

$$\beta_{ya.zw} = \frac{\sigma_{ya} - \sigma_{za}\sigma_{yz} - \sigma_{wy}\sigma_{wa}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \tag{11.35}$$

For Equation (11.19):

$$\begin{aligned}
ConfBias(Y, A) &= \beta_{ya} - \beta_{ya.zw} \\
&= \sigma_{ya} - \frac{\sigma_{ya} - \sigma_{za}\sigma_{yz} - \sigma_{wa}\sigma_{yw}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \\
&= \frac{\sigma_{ya}(1 - \sigma_{za}^2 - \sigma_{wa}^2) - \sigma_{ya} + \sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \\
&= \frac{\cancel{\sigma_{ya}} - \sigma_{ya}\sigma_{za}^2 - \sigma_{ya}\sigma_{wa}^2 - \cancel{\sigma_{ya}} + \sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw}}{1 - \sigma_{za}^2 - \sigma_{wa}^2} \\
&= \frac{\sigma_{za}\sigma_{yz} + \sigma_{wa}\sigma_{yw} - \sigma_{ya}(\sigma_{za}^2 + \sigma_{wa}^2)}{1 - \sigma_{za}^2 - \sigma_{wa}^2}
\end{aligned} \tag{11.36}$$

For Equation (11.20):

$$\begin{aligned}
ConfBias(Y, A) &= \beta_{ya} - \beta_{ya.zw} \\
&= \alpha + \beta\gamma + \lambda\delta - \frac{\alpha + \beta\gamma + \lambda\delta - \beta(\gamma + \beta\alpha) - \delta(\lambda + \delta\alpha)}{1 - \beta^2 - \delta^2} \\
&= \alpha + \beta\gamma + \lambda\delta - \frac{\alpha + \cancel{\beta\gamma} + \cancel{\lambda\delta} - \cancel{\beta\gamma} - \beta^2\alpha - \cancel{\delta\lambda} - \delta^2\alpha}{1 - \beta^2 - \delta^2} \\
&= \cancel{\alpha} + \beta\gamma + \lambda\delta - \frac{\cancel{\alpha(1 - \beta^2 - \delta^2)}}{\cancel{1 - \beta^2 - \delta^2}} \\
&= \beta\gamma + \lambda\delta
\end{aligned} \tag{11.37}$$

□

It is important to mention that although Theorem 11.5 assumes that the variables are standardized, the equations can be easily generalized to the non-standardized variables case. Moreover, the proof is general and can be extended to the case where the two confounders are not independent.

11.3 Selection bias

Selection bias occurs when there is a collider variable caused by both the sensitive attribute A and the outcome variable Y and the data generation process implicitly conditions on that collider variable. The simplest case is illustrated in Figure 11.7. Consistent with previous work, a box around a variable indicates that data is generated by conditioning on that variable.

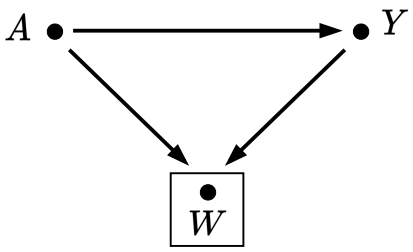


Figure 11.7: Simple collider structure

11.3.1 Binary Model

Definition 11.3.1. Given the basic collider structure (Figure 11.7), selection bias is defined as:

$$SelBias(Y, A, W) = StatDisp(Y, A)_W - StatDisp(Y, A) \tag{11.38}$$

Theorem 11.6. Assuming that $\mathbb{P}(a_0) = \mathbb{P}(a_1) = \frac{1}{2}$, the difference in discrimination due to selection bias is equal to:

$$SelBias(Y, A) = (1 - \mathbb{P}(w_0|a_0) - \mathbb{P}(w_1))(-\alpha + \beta - \gamma + \delta) \tag{11.39}$$

where α, β, γ, and δ denote, respectively, $\mathbb{P}(y_1|a_0, z_0), \mathbb{P}(y_1|a_0, z_1), \mathbb{P}(y_1|a_1, z_0)$, and $\mathbb{P}(y_1|a_1, z_1)$.

Proof. The proof is based on the proof of Theorem 11.2. Notice that, conditioning on variable Z in $ACE(Y, A)$ has the same formulation as conditioning on W in $StatDisp(Y, A)_W$. The difference is that the conditioning is on W instead of Z. The other important difference is that in Theorem 11.2, the unconditional expression $StatDisp(A, Y)$ is the biased estimation of the discrimination and the conditional expression $ACE(Y, A)$ is the unbiased estimation. Whereas in Theorem 11.6, it is the opposite: the unconditional expression $StatDisp(A, Y)$ is the unbiased estimation of discrimination and the conditional expression $StatDisp(Y, A)_W$ is the biased estimation. Hence, selection bias is just the opposite of Equation (11.11) while replacing the variable Z by the variable W. □

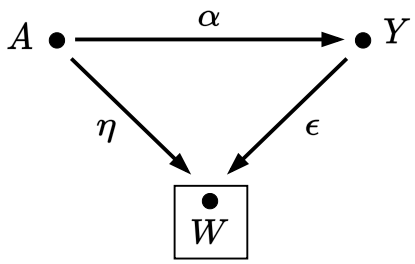


Figure 11.8: Simple collider structure with linear coefficients.

11.3.2 Linear Model Case

THEOREM 11.7. Let A, Y, and Z variables with linear regressions coefficients as in Figure 11.8. Bias due to selection is equal to:

$$SelBias(Y, A) = \frac{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_{wa}^2 - \sigma_{wa}\sigma_{yw}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} \tag{11.40}$$

Selection bias can also be expressed in terms of the linear regression coefficients as follows:

$$SelBias(Y, A) = \epsilon \ \frac{\sigma_a^4\alpha^2\eta + \sigma_a^4\alpha^3\epsilon - \sigma_y^2\sigma_a^2\eta - \sigma_y^2\sigma_a^2\alpha\epsilon}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \tag{11.41}$$

COROLLARY 11.8. For standardized variables A, Y, and W, selection bias can be expressed in terms of convariances as:

$$SelBias(Y, A) = \frac{\sigma_{ya}\sigma_{wa}^2 - \sigma_{wa}\sigma_{yw}}{1 - \sigma_{wa}^2} \tag{11.42}$$

And in terms of regression coefficient:

$$SelBias(Y, A) = \epsilon \ \frac{\alpha^2\eta + \alpha^3\epsilon - \eta - \alpha\epsilon}{1 - (\eta + \alpha\epsilon)^2} \tag{11.43}$$

Equations (11.42) and (11.43) can be obtained from Equations (11.40) and (11.41) as $\sigma_a = \sigma_w = \sigma_y = 1$.

Proof. For Equation (11.40),

$$\begin{aligned}
SelBias(Y, A) &= \beta_{ya.w} - \beta_{ya} \\
&= \frac{\sigma_w^2 \sigma_{ya} - \sigma_{yw}\sigma_{wa}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} - \frac{\sigma_{ya}}{\sigma_a^2} \\
&= \frac{(\sigma_w^2 \sigma_{ya} - \sigma_{yw}\sigma_{wa}) - \frac{\sigma_{ya}}{\sigma_a^2}(\sigma_a^2\sigma_w^2 - \sigma_{wa}^2)}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} \\
&= \frac{\cancel{\sigma_w^2\sigma_{ya}} - \sigma_{yw}\sigma_{wa} - \cancel{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_a^2\sigma_w^2} + \frac{\sigma_{ya}}{\sigma_a^2}\sigma_{wa}^2}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} \\
&= \frac{\frac{\sigma_{ya}}{\sigma_a^2}\sigma_{wa}^2 - \sigma_{wa}\sigma_{yw}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2}
\end{aligned}$$

For Equation (11.41),

$$\begin{aligned}
SelBias(Y, A) &= \beta_{ya.w} - \beta_{ya} \\
&= \frac{\sigma_w^2 \sigma_{ya} - \sigma_{yw}\sigma_{wa}}{\sigma_a^2\sigma_w^2 - \sigma_{wa}^2} - \frac{\sigma_{ya}}{\sigma_a^2} \\
&= \frac{\sigma_w^2\sigma_a^2\alpha - (\sigma_y^2\epsilon + \sigma_a^2\alpha\eta)(\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} - \frac{\cancel{\sigma_a^2}\alpha}{\cancel{\sigma_a^2}} \\
&= \frac{\sigma_w^2\sigma_a^2\alpha - \sigma_y^2\sigma_a^2\epsilon\eta - \sigma_y^2\sigma_a^2\alpha\epsilon^2 - \sigma_a^4\alpha\eta^2 - \sigma_a^4\alpha^2\eta\epsilon}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&\quad - \frac{\alpha(\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2)}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&= \frac{\cancel{\sigma_w^2\sigma_a^2\alpha} - \sigma_y^2\sigma_a^2\epsilon\eta - \sigma_y^2\sigma_a^2\alpha\epsilon^2 - \cancel{\sigma_a^4\alpha\eta^2} - \cancel{\sigma_a^4\alpha^2\eta\epsilon}}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&\quad + \frac{\cancel{-\sigma_a^2\sigma_w^2\alpha} + \cancel{\sigma_a^4\alpha\eta^2} + \cancel{2}\sigma_a^4\alpha^2\eta\epsilon + \sigma_a^4\alpha^3\epsilon^2}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2} \\
&= \epsilon\ \frac{\sigma_a^4\alpha^2\eta + \sigma_a^4\alpha^3\epsilon - \sigma_y^2\sigma_a^2\eta - \sigma_y^2\sigma_a^2\alpha\epsilon}{\sigma_a^2\sigma_w^2 - (\sigma_a^2\eta + \sigma_a^2\alpha\epsilon)^2}
\end{aligned}$$

□

11.4 Measurement bias

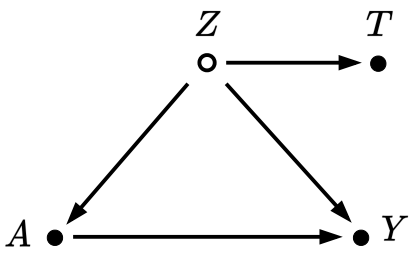


Figure 11.9: Simple measurement bias structure

Measurement bias arises from the way that particular variable(s) are measured. A common example is when the ideal variable for a model is not measurable/observable and instead we rely on a proxy variable which behaves differently in different groups. Figure 11.9 shows a simple scenario when measuring accurately discrimination based on A requires adjusting on variable Z. However, if Z is not measurable but a proxy variable T is measurable, measurement bias occurs when we adjust on T instead of Z.

Binary model

Definition 11.4.1. Given variables A, Y, Z, and T with causal relations as in Figure 11.9, measurement bias can be defined as:

$$MeasBias(Y, A) = StatDisp(Y, A)_T - StatDisp_Z(Y, A) \tag{11.44}$$

Theorem 11.9. Assuming that Z is not measurable, but only the error mechanism ($\mathbb{P}(T|Z)$) is available, and that $\mathbb{P}(a_0) = \mathbb{P}(a_1) = \frac{1}{2}$, the difference in discrimination due to measurement bias, $MeasBias(Y, A)$ can be expressed in terms of $\mathbb{P}(T|Z)$ as follows:

$$\begin{aligned}
&\epsilon(\delta - \beta) + (1 - \epsilon)(\gamma - \alpha) \\
&\quad - \epsilon\big(\delta - \beta + 4\mathbb{P}(t_1|z_0)(\beta - \delta + \gamma\Theta + \gamma\Psi)\big)\, Q \\
&\quad - (1 - \epsilon)\big(\gamma - \alpha + 4\mathbb{P}(t_0|z_1)(\alpha - \gamma + \delta + \delta\Psi^{-1} + \beta\Theta^{-1})\big)\, R
\end{aligned} \tag{11.45}$$

where:

$$\alpha = \mathbb{P}(y_1|a_0, t_0) \quad \gamma = \mathbb{P}(y_1|a_1, t_0) \quad Q = \frac{1 - \frac{\mathbb{P}(t_0|z_1)}{\epsilon}}{1 - \frac{\mathbb{P}(t_0|z_1)}{2\epsilon}} \quad \Phi = \frac{\epsilon + \frac{\tau}{2} - 1}{\epsilon + \frac{\tau}{2} - \frac{1}{2}} \quad \epsilon = \mathbb{P}(t_1)$$

$$\beta = \mathbb{P}(y_1|a_0, t_1) \quad \delta = \mathbb{P}(y_1|a_1, t_1) \quad R = \frac{1 - \frac{\mathbb{P}(t_1|z_0)}{1-\epsilon}}{1 - \frac{\mathbb{P}(t_1|z_0)}{2-2\epsilon}} \quad \Psi = \frac{1 - \tau}{\tau} \quad \tau = \mathbb{P}(t_0|a_0)$$

Proof. Let $\mathbb{P}(t_1) = \epsilon$ ($\epsilon \in]0, 1[$) and hence $\mathbb{P}(t_0) = 1 - \epsilon$. And let $\mathbb{P}(y_1|a_0, t_0) = \alpha$, $\mathbb{P}(y_1|a_0, t_1) = \beta$, $\mathbb{P}(y_1|a_1, t_0) = \gamma$, and $\mathbb{P}(y_1|a_1, t_1) = \delta$. Finally, let $\mathbb{P}(t_0|a_0) = \tau$. The remaining conditional probabilities of T given A are equal to the following:

$$\begin{aligned}
\mathbb{P}(t_1|a_0) &= 1 - \mathbb{P}(t_0|a_0) = 1 - \tau \\
\mathbb{P}(t_1|a_1) &= \frac{\mathbb{P}(t_1) - \mathbb{P}(t_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(a_1)} \\
&= 2\epsilon + \tau - 1 && (11.46) \\
\mathbb{P}(z_0|a_1) &= 1 - \mathbb{P}(z_1|a_1) && (11.47) \\
&= 2 - 2\epsilon - \tau
\end{aligned}$$

According to Definition 11.4.1:

$$MeasBias(Y, A) = StatDisp_T(Y, A) - StatDisp_Z(Y, A)$$

By the proof of Theorem 11.2, the first term:

$$StatDisp_T(Y, A) = \epsilon(\delta - \beta) + (1 - \epsilon)(\gamma - \alpha)$$

The rest of the proof consists in expressing $StatDisp_Z(Y, A)$ in terms of the error term $\mathbb{P}(T|Z)$.

$$StatDisp_Z(Y, A) = \mathbb{P}(y_1|do(a_1)) - \mathbb{P}(y_1|do(a_0)) \tag{11.48}$$

where:

$$\begin{aligned}
&\mathbb{P}(y_1|do(a)) = \\
&\frac{\mathbb{P}(y_1, a, t_1)}{\mathbb{P}(a|t_1)} \frac{\left(1 - \frac{\mathbb{P}(t_1|z_0)}{\mathbb{P}(t_1|a,y_1)}\right)\left(1 - \frac{\mathbb{P}(t_1|z_0)}{\mathbb{P}(t_1)}\right)}{1 - \mathbb{P}(t_1|z_0)\frac{\mathbb{P}(a)}{\mathbb{P}(t_1)}} \\
&+ \frac{\mathbb{P}(y_1, a, t_0)}{\mathbb{P}(a|t_0)} \frac{\left(1 - \frac{\mathbb{P}(t_0|z_1)}{\mathbb{P}(t_0|a,y_1)}\right)\left(1 - \frac{\mathbb{P}(t_0|z_1)}{\mathbb{P}(t_0)}\right)}{1 - \mathbb{P}(t_0|z_1)\frac{\mathbb{P}(a)}{\mathbb{P}(t_0)}}
\end{aligned} \tag{11.49}$$

The proof can be found in [229] (Section 3). Using Bayes rule, we can easily show that

$$\begin{aligned}
\mathbb{P}(y_1, a_1, t_1) &= \epsilon\delta + \frac{\delta\tau}{2} - \frac{\delta}{2} \\
\mathbb{P}(y_1, a_1, t_0) &= \gamma - \epsilon\gamma + \frac{\delta\tau}{2} - \frac{\tau\gamma}{2} \\
\mathbb{P}(y_1, a_0, t_1) &= \frac{\beta}{2} - \frac{\beta\tau}{2} \\
\mathbb{P}(y_1, a_0, t_0) &= \frac{\gamma\tau}{2}
\end{aligned}$$

Using Bayes rule and the marginal conditional probability rule: $\mathbb{P}(A|B) = \sum_{z \in Z} \mathbb{P}(A|B, z)\mathbb{P}(z|B)$, we can easily show that:

$$\begin{aligned}
\mathbb{P}(t_1|a_0, y_1) &= \frac{1}{4}\frac{\beta - \beta\tau}{\alpha\tau + \beta - \beta\tau} \\
\mathbb{P}(t_0|a_0, y_1) &= \frac{1}{4}\frac{\gamma\tau}{\gamma\tau + \beta - \beta\tau} \\
\mathbb{P}(t_1|a_1, y_1) &= \frac{\epsilon\delta + \frac{\delta\tau}{2} - \frac{\delta}{2}}{4\gamma - 4\epsilon\gamma - 2\tau\gamma + 4\epsilon\delta + 2\delta\tau - 2\delta} \\
\mathbb{P}(t_0|a_1, y_1) &= \frac{\gamma - \epsilon\gamma - \frac{\tau\gamma}{2}}{4\gamma - 4\epsilon\gamma - 2\tau\gamma + 4\epsilon\delta + 2\delta\tau - 2\delta}
\end{aligned}$$

Finally, using Bayes rule, we can show that:

$$\begin{aligned}
\mathbb{P}(a_0|t_1) &= \frac{\mathbb{P}(t_1|a_0)\mathbb{P}(a_0)}{\mathbb{P}(t_1)} = \frac{(1-\tau)}{2\epsilon} \\
\mathbb{P}(a_0|t_0) &= \frac{\mathbb{P}(t_0|a_0)\mathbb{P}(a_0)}{\mathbb{P}(t_0)} = \frac{\tau}{2-2\epsilon} \\
\mathbb{P}(a_1|t_1) &= \frac{\mathbb{P}(t_1|a_1)\mathbb{P}(a_1)}{\mathbb{P}(t_1)} = \frac{\epsilon + \frac{\tau}{2} - \frac{1}{2}}{\epsilon} \\
\mathbb{P}(a_1|t_0) &= \frac{\mathbb{P}(t_0|a_1)\mathbb{P}(a_1)}{\mathbb{P}(t_0)} = \frac{(2-2\epsilon-\tau}{2-2\epsilon}
\end{aligned}$$

After some algebra, we have:

$$\begin{aligned}
ACE(Y,A) = \epsilon\big(\delta - \beta + 4\mathbb{P}(t_1|z_0)(\beta - \delta + \gamma\Phi + \gamma\Psi)\big)\ Q \\
+ (1-\epsilon)\big(\gamma - \alpha + 4\mathbb{P}(t_0|z_1)(\alpha - \gamma + \delta + \delta\Psi^{-1} + \beta\Phi^{-1})\big)\ R
\end{aligned} \tag{11.50}$$

□

Linear Model Case

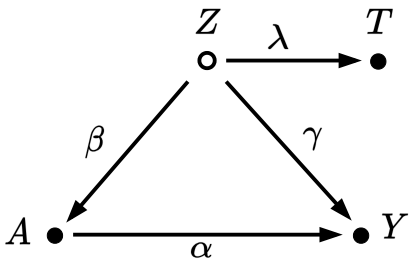


Figure 11.10: Simple measurement bias structure with linear coefficients.

THEOREM 11.10. Let A, Y, Z, and T variables with linear regressions coefficients as in Figure 11.10 which represents the basic measurement bias structure. Bias due to measurement error is equal to:

$$MeasBias(Y,A) = \frac{{\sigma_z}^2\beta\gamma({\sigma_t}^2 - {\sigma_z}^2\lambda^2)}{{\sigma_a}^2{\sigma_t}^2 - {\sigma_z}^4\lambda^2\beta^2} \tag{11.51}$$

COROLLARY 11.11. For standardized variables A, Y, Z, and T, measurement bias is equal to:

$$MeasBias(Y,A) = \frac{\beta\gamma(1-\lambda^2)}{1-\lambda^2\beta^2} \tag{11.52}$$

Proof.

$$
\begin{aligned}
MeasBias(Y, A) &= \beta_{ya.t} - \beta_{ya.z} \\
&= \frac{\sigma_t^2\sigma_{ya} - \sigma_{yt}\sigma_{ta}}{\sigma_a^2\sigma_t^2 - \sigma_{ta}^2} - \frac{\sigma_z^2\sigma_{ya} - \sigma_{yz}\sigma_{za}}{\sigma_a^2\sigma_z^2 - \sigma_{za}^2} \\
&= \frac{\sigma_t^2(\sigma_a^2\alpha + \sigma_z^2\beta\gamma) - (\sigma_z^2\gamma\lambda + \sigma_z^2\alpha\beta\lambda)(\sigma_z^2\lambda\beta)}{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2} - \alpha \qquad (11.53) \\
&= \frac{\sigma_t^2\sigma_a^2\alpha + \sigma_t^2\sigma_z^2\beta\gamma - \sigma_z^4\gamma\lambda^2\beta - \sigma_z^4\lambda^2\beta^2\alpha}{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2} \\
&= \frac{\cancel{\alpha}\,(\cancel{\sigma_t^2\sigma_a^2 - \sigma_z^2\lambda^2\beta^2})}{\cancel{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2}} + \frac{\sigma_t^2\sigma_z^2\beta\gamma - \sigma_z^4\gamma\lambda^2\beta}{\sigma_a^2\sigma_t^2 - \sigma_z^4\lambda^2\beta^2} - \cancel{\alpha} \\
&= \frac{{\sigma_z}^2\beta\gamma({\sigma_t}^2 - {\sigma_z}^2\lambda^2)}{{\sigma_a}^2{\sigma_t}^2 - {\sigma_z}^4\lambda^2\beta^2}
\end{aligned}
$$

In step (11.53), $\beta_{ya.z}$ is replaced by α (see proof of Theorem 11.3). □

11.5 Interaction bias

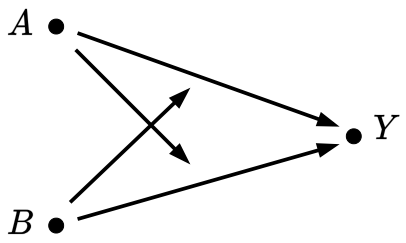


Figure 11.11: Interaction Bias, where A and B are sensitive variables and Y is an outcome.

Interaction bias takes place in the presence of two sensitive attributes when the value of one sensitive attribute influences the effect of the other sensitive attribute on the outcome. Interaction bias is graphically illustrated in Figure 11.11. Note that regular DAGs are not able to express interaction. For this reason, we are employing the graphical representation proposed by [230]. The arrows pointing to arrows, instead of nodes account for the interaction term. In a binary model, interaction bias coincides with interaction term ($Interaction$) in the case of an intersectional sensitive attribute. Interaction bias also affects the individual measurement of the effect of sensitive attribute A or B.

11.5.1 Binary model, Intersectional Sensitive Variable

Given binary sensitive variables A, B and a binary outcome Y, the joint discrimination of $A = 0$ and $B = 0$ with respect to Y can be defined as follows:

DEFINITION 11.5.1.

$$StatDisp(Y, A, B) = P(Y_1|a_1, b_1) - P(Y_1|a_0, b_0) \qquad (11.54)$$

Here $Y = 1$ is a positive outcome, $A = 1$ and $B = 1$ (a_1, b_1) represent the disadvantaged group.

THEOREM 11.12. Under the assumption of no common parent for A and Y and B and Y [5] we can express $StatDisp(Y, A, B)$ in terms of causal effects of A and B and interaction between A and B on the additive scale:

$$\begin{aligned} StatDisp(Y, A, B) = &\big[P(Y_1|a_1, b_0) - P(Y_1|a_0, b_0)\big] + \big[P(Y_1|a_0, b_1) - P(Y_1|a_0, b_0)\big] \\ &+ Interaction(A, B) \end{aligned}$$

where $Interaction(A, B) = P(Y_1|a_1, b_1) - P(Y_1|a_0, b_1) - P(Y_1|a_1, b_0) + P(Y_1|a_0, b_0)$

Proof. By Definition 11.5.1:

$$\begin{aligned} StatDisp(Y, A, B) &= P(Y_1|a_1, b_1) - P(Y_1|a_0, b_0) \\ &= P(Y_1|a_1, b_1) - P(Y_1|a_0, b_0) + P(Y_1|a_0, b_0) - P(Y_1|a_0, b_0) \\ &\quad + P(Y_1|a_1, b_0) - P(Y_1|a_1, b_0) + P(Y_1|a_0, b_1) - P(Y_1|a_0, b_1) \\ &= P(Y_1|a_1, b_0) - P(Y_1|a_0, b_0) + P(Y_1|a_0, b_1) - P(Y_1|a_0, b_0) \\ &\quad + P(Y_1|a_1, b_1) - P(Y_1|a_0, b_1) - P(Y_1|a_1, b_0) + P(Y_1|a_0, b_0) \\ &= \big[P(Y_1|a_1, b_0) - P(Y_1|a_0, b_0)\big] + \big[P(Y_1|a_0, b_1) - P(Y_1|a_0, b_0)\big] \\ &\quad + Interaction(A, B) \end{aligned}$$

□

Notice that: $P(Y_1|a_1, b_0) - P(Y_1|a_0, b_0)$ is the effect of A on Y in case there is no interaction, and similarly for B: $P(Y_1|a_0, b_1) - P(Y_1|a_0, b_0)$ is the effect of B on Y in case there is no interaction. To avoid confusion, we denote such expressions as $SD_{\not{Int}}(Y, A)$ and $SD_{\not{Int}}(Y, B)$ respectively.

DEFINITION 11.5.2. Under the assumption of no confounders between A and Y on one hand, and between B and Y on the other hand, adding up the single effects of A and B on Y to estimate the discrimination due to both sensitive variables $StatDisp(Y, A, B)$ leads to a biased estimation. The amount of the bias ($StatDisp$) coincides with the interaction term as follows:

$$\begin{aligned} IntBias(Y, A, B) &= StatDisp(Y, A, B) \\ &\quad - \big[SD_{\not{Int}}(Y, A) + SD_{\not{Int}}(Y, B)\big] \\ &= Interaction(A, B) \end{aligned}$$

The presence of the $Interaction(A, B)$ in the case of two sensitive variables is very common.

[5] This assumption is relatively easy to satisfy in case of immutable sensitive attributes such as gender or race because they are unlikely to have external causes. It is important to control for possible confounders when sensitive attributes can have external causes, for example, political beliefs can be influenced by education.

[231] distinguish 16 combinations of the effects of binary A and B on the binary outcome Y. Only six of those cases correspond to $Interaction(A, B) = 0$, which means that there is no interaction. Indeed, the interaction is absent only when at least one of the terms A and B has no effect on Y [231]. Unfortunately, most of the time the numeric value of interaction does not indicate which particular case (out of 16 combinations of the effects of A and B) is dominant in the data. However, VanderWeele and Robins show that under sufficient-component-cause framework and assumption of monotonic effect of A and B [6] if $P(Y_1|a_1, b_1) - P(Y_1|a_0, b_1) - P(Y_1|a_1, b_0) > 0$, then the synergism between $A = 1$ and $B = 1$ must be present [231, 232]. In the fairness scenario, this means that two privileged groups have a synergetic effect on the positive outcome. In terms of the previous example, it is a situation where only conservative men are hired.

11.5.2 Binary model, Individual Sensitive Variable

Given binary sensitive variables A, B and a binary outcome Y, the discrimination with respect to only A (and similarly for B) can be expressed as follows:

Definition 11.5.3.

$$StatDisp(Y, A) = P(Y_1|a_1) - P(Y_1|a_0) \tag{11.55}$$

Theorem 11.13. Under the previously introduced assumption of no confounding, discrimination with respect to A can be decomposed into an interaction-free discrimination and the interaction between A and B:

$$StatDisp(Y, A) = SD_{\not{Int}}(Y, A) + P(b_1)Interaction(A, B)$$

Proof.

$$\begin{aligned}
StatDisp(Y, A) &= \mathbb{P}(Y_1|a_1) - \mathbb{P}(Y_1|a_0) \\
&= \sum_b \mathbb{P}(Y_1|a_1, b)\mathbb{P}(b|a_1) - \sum_b \mathbb{P}(Y_1|a_0, b)\mathbb{P}(b|a_0) \\
&= \mathbb{P}(Y_1|a_1, b_1)\mathbb{P}(b_1|a_1) + \mathbb{P}(Y_1|a_1, b_0)\mathbb{P}(b_0|a_1) \\
&\quad - \mathbb{P}(Y_1|a_0, b_1)\mathbb{P}(b_1|a_0) - \mathbb{P}(Y_1|a_0, b_0)\mathbb{P}(b_0|a_0) \\
&= \mathbb{P}(Y_1|a_1, b_1)\mathbb{P}(b_1|a_1) + \mathbb{P}(Y_1|a_1, b_0)\mathbb{P}(1 - \mathbb{P}(b_1|a_1) \\
&\quad - \mathbb{P}(Y_1|a_0, b_1)\mathbb{P}(b_1|a_0) - \mathbb{P}(Y_1|a_0, b_0)\mathbb{P}(1 - \mathbb{P}(b_1|a_0) \\
&= \mathbb{P}(b_1|a_1)\big(\mathbb{P}(Y_1|a_1, b_1) - \mathbb{P}(Y_1|a_1, b_0)\big) + \mathbb{P}(Y_1|a_1, b_0) \\
&\quad + \mathbb{P}(b_1|a_0)\big(\mathbb{P}(Y_1|a_0, b_0) - \mathbb{P}(Y_1|a_0, b_1)\big) - \mathbb{P}(Y_1|a_0, b_0)
\end{aligned}$$

[6] monotonic effect means, that an intervention either increases or decreases outcome Y for every individual.

Since A and B are independent $\mathbb{P}(b_1|a_1) = \mathbb{P}(b_1|a_0) = \mathbb{P}(b_1)$. It follows that:

$$\begin{aligned} StatDisp(Y, A) &= \mathbb{P}(b_1)Int(A, B) + \mathbb{P}(Y_1|a_1, b_0) - \mathbb{P}(Y_1|a_0, b_0) \\ &= \mathbb{P}(b_1)Int(A, B) + SD_{\not{Int}}(Y, A) \end{aligned}$$

□

$StatDisp(Y, B)$ can be decomposed in a similar way. Interaction bias $IntBias(Y, A)$ can then be defined as:

DEFINITION 11.5.4.

$$\begin{aligned} IntBias(Y, A) &= StatDisp(Y, A) - SD_{\not{Int}}(Y, A) \\ &= P(b_1)Interaction(A, B) \end{aligned}$$

$IntBias(Y, B)$ can be defined similarly:

$$\begin{aligned} IntBias(Y, B) &= StatDisp(Y, B) - SD_{\not{Int}}(Y, B) \\ &= P(a_1)Interaction(A, B) \end{aligned}$$

11.5.3 Linear Model Case

Given the true model:

$$Y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 AB + \beta_4 C \tag{11.56}$$

And biased model, that does not include interaction term β_3:

$$Y = \beta_0' + \beta_1' A + \beta_2' B + \beta_4' C \tag{11.57}$$

Where A and B are binary sensitive attributes, C is a set of covariates and Y is a continuous outcome (for example a credit score).

The change in Y due to A is $\beta_1 + \beta_3 B$ and, similarly the change in Y due to B is $\beta_2 + \beta_3 A$ [233]. In this case, a measure of effect of A (β_1') or B (β_2') without an interaction term would be inaccurate. Next we define the bias introduced by not accounting for the interaction between two sensitive attributes.

Linear model, Intersectional Sensitive Variable

THEOREM 11.14. Let A,B and Y be variables with linear regression coefficients as in Equation 11.57. In a linear model with binary A and B the bias due to interaction, when measuring the effect of intersectional sensitive variable A and B

on Y ($StatDisp(Y, A, B)$) is equal to:

$$\begin{aligned} IntBias(Y, A, B) &= (\beta_1^{'} + \beta_2^{'}) - (\beta_1 + \beta_2) \\ &= \beta_3 \end{aligned}$$

Proof. $\beta_1^{'} + \beta_2^{'}$ represents the causal effect of A and B on Y including interaction, whereas $\beta_1 + \beta_2$ represents the same causal effect but without interaction. The difference coincides with the interaction coefficient β_3. □

Intuitively, β_3 is part of an effect of the intersectional sensitive variable $A = 1$, $B = 1$ on Y that is left out of the estimation when fitting linear regression without the interaction term.

Linear model, Individual Sensitive Variable

The difference of measurement of effect of A on Y with interaction term ($\beta_1^{'}$) and without interaction term (β_1) depends on the value of B.

> **THEOREM 11.15.** Let A,B and Y be variables with linear regression coefficients as in Equation 11.57. In a linear model with binary A and B the bias due to interaction, when measuring the effect of A on Y ($StatDisp(Y, A)$) is equal to:
>
> $$\begin{aligned} IntBias(Y, A) &= \beta_1^{'} - \beta_1 \\ &= \beta_3 \mathbb{P}(B_1) \end{aligned}$$

Proof. $StatDisp(Y, A)$ measures how wrong is the evaluation of effect of $A = 1$ on average, for cases where $B = 1$ or $B = 0$, which are as follows:

$$\beta_1^{'} = \begin{cases} \beta_1 + \beta_3 & \text{when} \quad B = 1 \\ \beta_1 & \text{when} \quad B = 0 \end{cases} \tag{11.58}$$

Note that the $StatDisp(Y, A)$ is dependent on β_3 and the probability of $B = 1$ (and, similarly, $StatDisp(B, A)$ is dependent on β_3 and the probability of $A = 1$). □

11.6 Bias analysis

Expressing different types of bias in terms of the model parameters (conditional probabilities and regression coefficients) allows to study the behavior of bias and how it is impacted by the different parameters. In particular, at which parameters value it is peaked and at which other values it is absent. The aim is to identify the cases where a given estimation of discrimination is biased and at which extent.

11.6.1 Binary Case

Confounder Bias is absent when at least one of the two terms of Equation 11.11 is equal to 0. For the first term $(1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1) = 0)$, it is easy to show that it is equivalent to $\mathbb{P}(z_0|a_1) = \mathbb{P}(z_0)$ which in turn means that Z and A are independent ($A \perp\!\!\!\perp Z$).

The second term is equal to 0 when :

$$\mathbb{P}(y_1|a_0, z_0) - \mathbb{P}(y_1|a_0, z_1) = -(\mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_1, z_1)) \tag{11.59}$$

Thre right-hand side can be interpreted as the Contolled Direct Effect (CDE) [234] of Z on Y when $A = 0$ whereas the left-hand side is the opposite of $\mathbb{P}(y_1|a_1, z_0) - \mathbb{P}(y_1|a_1, z_1)$ which is the CDE of Z on Y when $A = 1$. Confounding bias is equal zero, when the CDE of Z on Y when $A = 1$ is the exact opposite of to that when $A = 0$. In the job hiring example of Figure 11.1(a), it means that we privilege poor liberals as much as we privilege rich conservatives, therefore the effect $Z- > Y$ is canceled out. Equation 11.59 can also hold when both sides are equal to 0. This means that Z has no direct effect on Y (no edge between Z and Y). Z can still have effect on Y which is mediated through A, but it does not have a role as a confounder. To summarize, confounding bias is absent in three cases: either $A \perp\!\!\!\perp Z$ (A and Z are independent) or the edge $Z \rightarrow Y$ is absent, or the CDE of Z on Y when $A = 0$ and $A = 1$ are opposite and hence cancel each others.

Confounding bias is peaked when the first term $(1 - \mathbb{P}(z_0|a_0) - \mathbb{P}(z_1))$ is equal to 1 or -1 *and* the second term $(-\alpha + \beta - \gamma + \delta)$ is equal to 2 or -2. The first term is equal to 1 when $\mathbb{P}(z_1) = 0$ and $\mathbb{P}(z_0|a_0) = 0$. This is an extreme situation when all data instances have the same values of A and Z variables, that is, a_1 and z_0. The same term is equal to -1 when $\mathbb{P}(z_1) = 1$ and $\mathbb{P}(z_0|a_0) = 1$ which corresponds to the other extreme situation of all data instances have a_0 and z_0. In the job hiring example, both cases correspond to a situation when all candidates are of the same type: poor liberals or rich liberals. The second term reaches a peak value (2.0 or -2.0) when the CDE of Z on Y is maximum (1 or -1) for both a_0 and a_1. To summarize, confounding bias is optimal when the effect through the edge $Z \rightarrow A$ is very strong (first term) *and* the effect through the edge $Z \rightarrow Y$ is very strong (second term). This optimal situation can be seen as an extreme case of Simpson's paradox [235].

Collider Bias Collider bias can be viewed as an inverse case of a confounder bias. While confounder bias compromises internal validity, selection bias is a threat to external validity [236]. Similarly, as confounder bias, collider bias does not manifest if the direct link between A and W or Y and W is absent, or the link between W and Y is the opposite for the values $A = 1$ and $A = 0$. The bias is maximized when the group corresponding to $A = 1$ and $W = 0$ is very large (the negative bias case would occur if the group $A = 1$ and $W = 1$ is dominant). Maximization of bias also requires that the link from Y to W is deterministic and has the same direction for both values of A.

Measurement Bias depends heavily on $\mathbb{P}(T|Z)$. For instance, from Theorem 11.9, it is easy to show that if $\mathbb{P}(t_0|z_1) = \mathbb{P}(t_1|z_0) = 0$ (T and Z are fully dependent), then $Q = R = 1$, and consequently the measurement bias disappears. Conversely, if $\mathbb{P}(t_0|z_1) = \mathbb{P}(t_1) = \epsilon$ and $\mathbb{P}(t_1|z_0) = \mathbb{P}(t_0) = 1 - \epsilon$ (T and Z are independent), then $Q = R = 0$, and consequently, measurement bias

is maximized as the two negative terms of Equation (11.45) disappear. The maximum value of measurement bias in that case is $\epsilon(\delta - \beta) + (1 - \epsilon)(\gamma - \alpha)$.

Interaction Bias Interaction bias for the intersectional case coincides with the interaction term. More precisely, it is maximized when the interaction is maximized and diminishes when the interaction is small. Note that the interaction is equal 0 when one of the sensitive attributes does not have an effect on Y [231]. The interaction bias when measuring the effect of one sensitive attribute A or B on Y depends on the interaction term and the probability of $B = 1$ and $A = 1$, respectively. The bias increases with the probability of $A = 1$ or $B = 1$ and the interaction term. Interaction bias is equal to zero when either interaction, to the probability of $B = 1$ or $A = 1$, respectively, is equal to 0.

11.6.2 Linear Case

To analyze the different types of bias in the linear case, we generate synthetic data according to the following models. Without loss of generality, the range of possible values of all coefficients ($\alpha, \beta, \gamma, \eta, \epsilon$, and δ) is $[-1.0, 1.0]$:

Confounding Structure:

$$Z = \mathcal{U}_z,$$
$$A = \beta Z + \mathcal{U}_a,$$
$$Y = \alpha A + \gamma Z + \mathcal{U}_y$$

Colliding Structure:

$$A = \mathcal{U}_a,$$
$$Y = \alpha A + \mathcal{U}_y,$$
$$W = \eta A + \epsilon Y + \mathcal{U}_w$$

Measurement Structure:

$$Z = \mathcal{U}_z,$$
$$A = \beta Z + \mathcal{U}_a,$$
$$Y = \alpha A + \gamma Z + \mathcal{U}_y,$$
$$T = \delta Z + \mathcal{U}_t$$

$$\mathcal{U}_z \sim \mathcal{N}(0, 1),$$
$$\mathcal{U}_a \sim \mathcal{N}(0, 1),$$
$$\mathcal{U}_y \sim \mathcal{N}(0, 1),$$
$$\mathcal{U}_w \sim \mathcal{N}(0, 1),$$
$$\mathcal{U}_t \sim \mathcal{N}(0, 1).$$

Figure 11.14 shows the magnitude of each type of bias based on the expressions obtained in Sections 11.2, 11.3, and 11.4. In particular, Equations 11.16 for confounding bias, 11.41 for selection bias, and 11.51 for measurement bias. Three dimensions plot is used for confounding bias (Figure 11.14(a)) as bias is expressed in terms of two variables (β and γ) whereas four dimensions plots are used for selection and measurement biases (three variables). Confounding bias is maximized when both β and γ have extreme values (+1.0 or −1.0): positive bias when β and γ are of the same sign, and negative otherwise. Bias is absent when at least one of the coefficients is zero. In between these extreme cases, confounding bias has strictly linear relation with β whereas a non-linear relation with γ. More importantly, confounding bias is more sensitive to β than to γ particularly for extreme values (when coefficients are close to +1.0 or −1.0). That is, modifying the effect of the confounder (e.g. Z) on the sensitive variable (e.g. A) has more impact on the confounding bias than modifying the effect of the confounder on the outcome variable (e.g. Y) with the same amount. In the job hiring example (Section 10.2.1) this means that the effect of Socio-Economic status on politicial belief has more impact on the counfounding bias than the effect of socio-economic status on job hiring. However, if the variables are standardized, both effects contribute equally to confounding bias (Corollary 11.4).

Unlike confounding bias, the magnitude of selection bias (Figure 11.14(b)) depends also on the regression coefficient of Y on A (α). Selection bias is peaked in two cases depending on the value of α. First, when η and ϵ have the same extreme values (1 or −1) and $\alpha = 1$. This

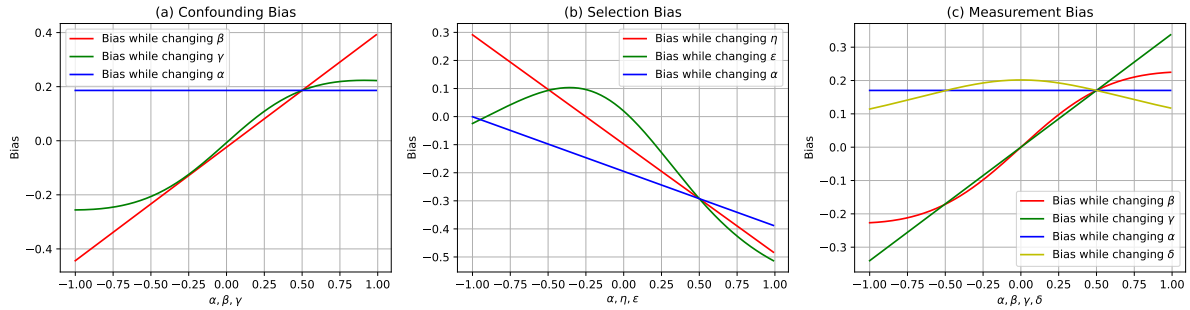


Figure 11.12: Bias Magnitude while changing one variable and holding the other variables at 0.5.

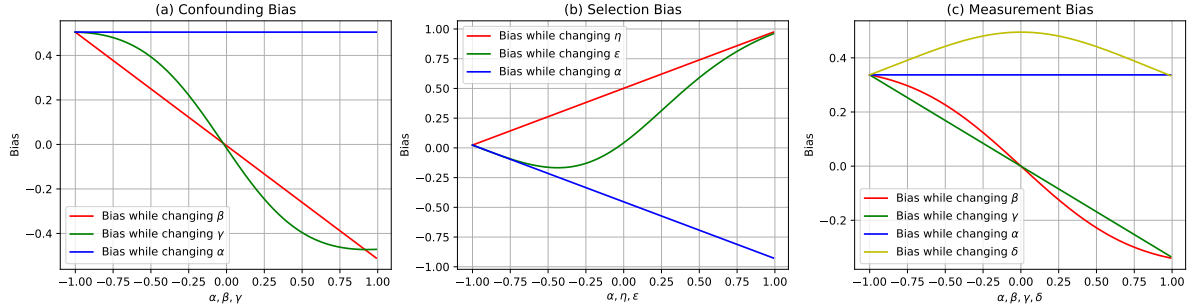


Figure 11.13: Bias Magnitude while changing one variable and holding the other variables at −1.0.

leads to maximal negative bias. Second, when η and ϵ have extreme but different sign values (1 or −1) and $\alpha = -1$. This corresponds to maximal positive bias. Intuitively, conditioning on the collider variable W introduces a spurious effect between the two causes A on Y: any information "explaining away" one cause will make the other cause more plausible. Using the job hiring example (Figure 11.1(b)), if there is maximum negative discrimination based on the political beliefs of the candidates ($\alpha = -1$) and we measure discrimination using only labor union records, while political belief and job hiring have strong but opposite effects on labor union membership, the selection bias will be maximum to the point it cancels out all positive discrimination and leads to a conclusion of no discrimination. Figure 11.14(b) shows also that selection bias disappears when ϵ is zero, but not when η is zero. When $\epsilon \neq 0$, selection bias can be zero depending on the value of ϵ as follows: $\epsilon = 1$ and $\alpha = -\eta$ or $\epsilon = -1$ and $\alpha = \eta$. In general, selection bias has a linear relationship with both α and η, while a non-linear relationship with ϵ[7].

Similarly to confounding and selection, the measurement bias (Figure 11.14(c)) is peaked when β and γ have extreme values (1.0 or −1.0) but when $\delta = 0$. This is expected as, by definition, the more Z and T are independent, the higher measurement bias is. Conversely, the plot shows that the measurement bias disappears as δ departs from 0[8].

[7] Such relations can be observed more clearly using 2D plots (Figures 11.12 and Figure 11.13).
[8] The 2D plots in the appendix show clearly these observations.

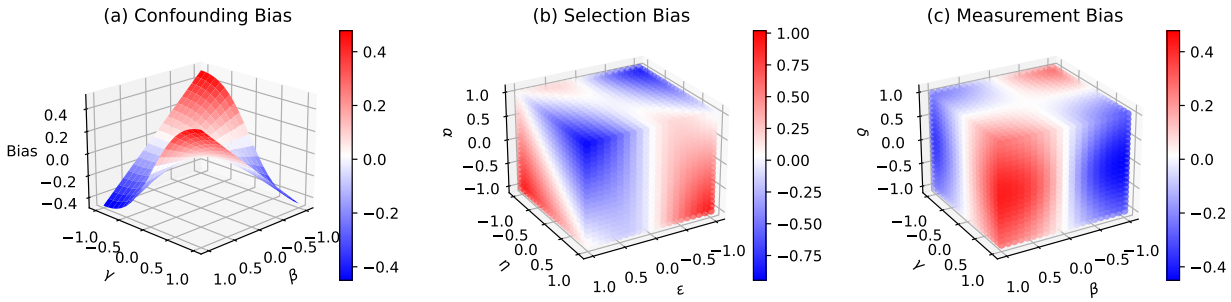


Figure 11.14: Bias magnitude in the linear case

11.7 Bias analysis in benchmark datasets

We use well-known fairness benchmark data sets [237] for the experiments on real data: Adult [9], Boston housing [10], Compas [238], Communities and crimes [11] and Dutch census [12] data. The causal experiments on the real data are limited by the availability of true causal graphs for the benchmark fairness datasets. Furthemore, [223] shows, that obtaining reliable causal graphs with causal discovery algorithms is a complicated task. However, we assume that the graphs in the literature are true for a given real dataset. We use the graphs by [239, 240] for Adult and Dutch data sets to measure the interaction bias. For measuring confounder and collider biases we rely on graphs obtained by [223] for Communities and Crimes (11.15), Boston Housing (11.16), Compas (11.18), and Dutch datasets (11.17).

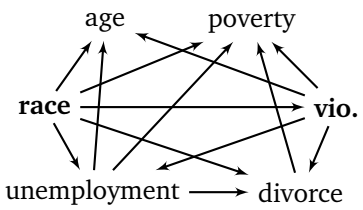


Figure 11.15: The graph for the communities and crime dataset. 'divorce', 'age', 'poverty' and 'unemployement' are the colliders between 'race' and 'violence' (vio.). The graph is produced using LiNGAM algorithm.

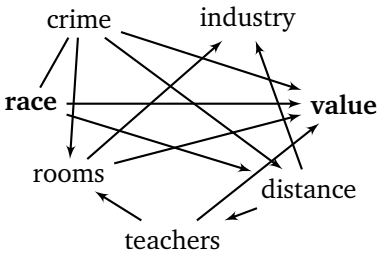


Figure 11.16: The graph for the Boston housing data set. 'Crime' is a possible confounder between 'race' and 'value'.The graph is produced using GES algorithm.

[9] https://archive.ics.uci.edu/dataset/2/adult
[10] http://lib.stat.cmu.edu/datasets/boston
[11] https://archive.ics.uci.edu/dataset/183/communities+and+crime
[12] https://microdata.worldbank.org/index.php/catalog/2102/data-dictionary

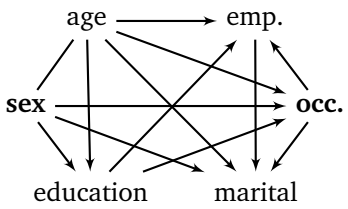


Figure 11.17: The graph for the Dutch data set. 'Marital Status' is a collider between 'sex' and 'occupation' (occ.). The graph is produced using GES algorithm.

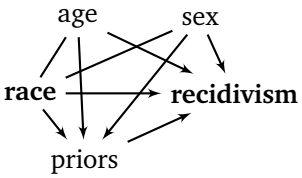


Figure 11.18: The graph for the Compas dataset. 'Age' and 'sex' are possible confounders between 'race' and 'recidivism'. The graph is produced using PC algorithm.

For measurement bias, we use synthetic data because the required structure is not present in the available graphs for the benchmark data sets. Synthetic data is generated according to the following models:

The variables Z, A, T, and Y are binary Bernoulli variables controlled by the parameter p_1. Conditional dependencies of the measurement bias structure define how the parameter p_1 depends on the value of the parent variables.

$$Z \sim (p) = \begin{cases} p_1, \\ p_0 = 1 - p_1 \end{cases} \qquad Y \sim (p; Z, A) = \begin{cases} p_1 = 0.5 * z + 0.5 * a, \\ p_0 = 1 - p_1 \end{cases}$$

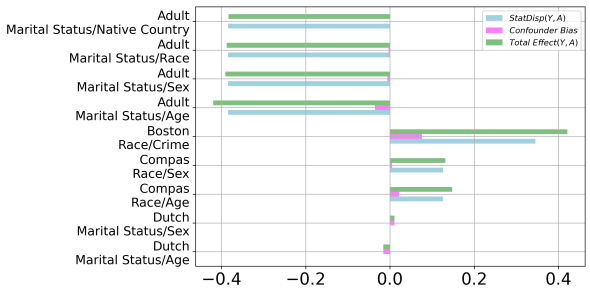
$$T \sim (Z; p) = \begin{cases} p_1, & \text{if } Z = 1, \\ p_0 = 1 - p_1 & \\ p_1^{'}, & \text{if } Z = 0. \\ p_0^{'} = 1 - p_1^{'} & \end{cases} \qquad A \sim (Z; p) = \begin{cases} p_1, & \text{if } Z = 1, \\ p_0 = 1 - p_1 & \\ p_1^{'}, & \text{if } Z = 0. \\ p_0^{'} = 1 - p_1^{'} & \end{cases}$$

The parameters p_1, p_0, $p_1^{'}$ and $p_0^{'}$ are generated randomly and take values between 0 and 1.

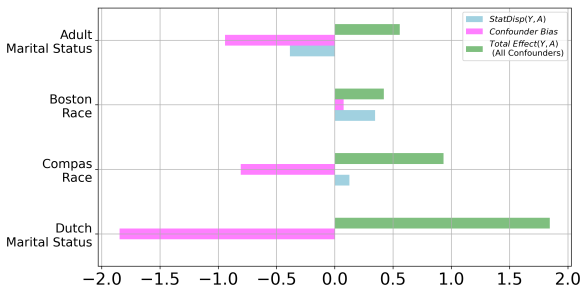
Although we cannot claim that the causal structure that we use for the experiments is the ground truth, it is useful for experimentally demonstrating the behavior of causal biases. In addition, the considered causal structures most often show the presence of multiple causal biases at once. However, for the purposes of illustration, we control for a single type of bias separately. More precisely, we consider the difference in measured discrimination with the presence of the absence of a certain type of bias.

The experimental results for confounder bias show that the biases for each individual confounding variable are not significant (Figure 11.19(a)). However, its magnitude increases and

can cancel out the value for statistical disparity (Dutch data set), when multiple confounders are considered simultaneously (Figure 11.19(b)). Measurement bias takes the highest value for *Synthetic2* dataset (Figure 11.20(b)). The effect of A on Y when controlling for T appears smaller than when controlling for Z. Here, the value of T is highly dependent on Z if $Z = 0$, but only loosely dependent on Z if $Z = 1$. The prior probability of Z conditions it to take value $Z = 1$ with probability 0.95. Therefore, the link between Z and T is weak. The weak link between the variables makes T a bad predictor for Z and introduces a high measurement bias. Collider bias (Figure 11.20(a)) is significant if it was introduced by conditioning on income (adult data), age (Compas data), economic status (Dutch data), poverty, unemployment, or divorce (Communities and crime data). Collider bias would reverse the value of statistical disparity, showing discrimination against the privileged group instead of discrimination against the disadvantaged group. We observe a portion of the interaction in all cases of the intersectional sensitive attribute (Figure 11.21(a)). However, the value of synergism is negative, which means that it is not present in the data. Measurement of interaction bias for A and B individually can yield different values of interaction bias (Figure 11.21(b)). Although the interaction term is symmetric for A and B, the interaction bias value is also dependent on the probability $B = 1$ (when measuring *IntBias*(Y, A)) or $A = 1$ (when measuring *IntBias*(Y, B)). Therefore, for example, the interaction bias for sex is higher than for age in the Adult data set, because the probability of value 1 for age is higher than the probability of the sex variable taking value 1. Furthermore, we observe that the statistical disparity does not always correspond to the sum of interaction bias and statistical disparity without interaction ($StatDisp(Y, A) \neq SD_{\not{Int}}(Y, A) + P(b_1)Interaction(A, B)$), as required in Theorem 11.13. This observation suggests that the two sensitive variables A and B are not independent as suggested by the graphs provided by [239, 240]. Indeed, the graphs discovered by [223] show the dependency between age and sex variables in the Dutch data set (Figure 11.17).

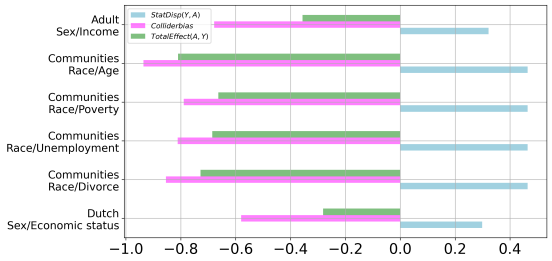


((a)) Confounder bias, when treating each confounder separately.

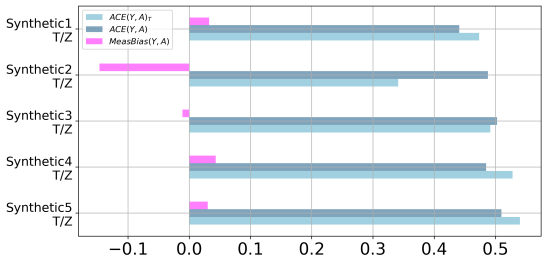


((b)) Confounder bias when treating all confounders together.

Figure 11.19: Confounder bias.

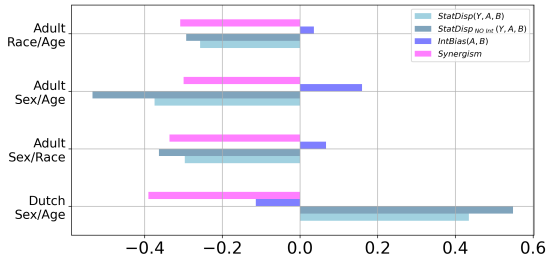


((a)) Collider bias, when treating each confounder separately.

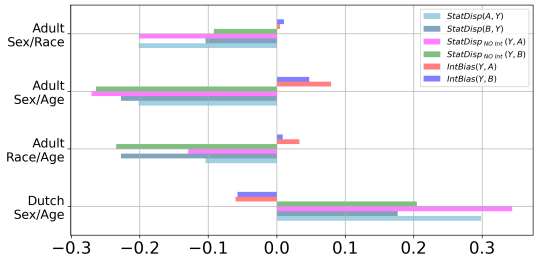


((b)) Measurement bias. Experiments on synthetic data sets.

Figure 11.20: Collider and measurement bias



((a)) Interaction bias, intersectional sensitive variable.



((b)) Interaction bias for individual sensitive attributes.

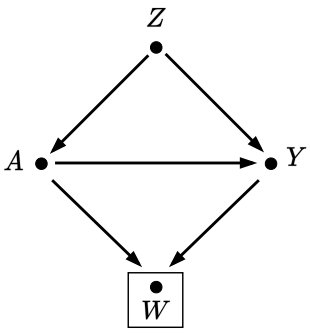
Figure 11.21: Interaction bias.

11.8 Concurrent biases

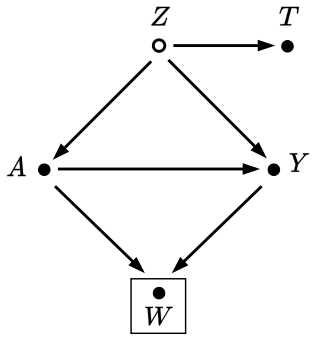
Confounding and selection biases. In presence of one or several confounder and collider variables, the estimation of discrimination can suffer from both confounding and selection biases simultaneously. Figure 11.22(a) shows the simplest case. According to Definitions 11.2.1 and 11.3.1, confounding bias can be isolated by adjusting on the confounder variable $ConfBias(Y,A) = StatDisp(Y,A) - StatDisp_Z(Y,A)$[13] ($\beta_{ya} - \beta_{ya.z}$ in the linear case), whereas selection bias can be isolated by cancelling the adjustment on the collider variable $SelBias(Y,A) = StatDisp_W(Y,A) - StatDisp(Y,A)$ ($\beta_{ya.w} - \beta_{ya}$ in the linear case). The total bias in presence of both types of bias can then be estimated as $StatDisp_W(Y,A) - StatDisp_Z(Y,A)$ in the binary case and $\beta_{ya.w} - \beta_{ya.z}$ in the linear case.

Confounding and measurement biases. Measurement bias (Figure 11.9) is defined as the difference in estimating *StatDisp* when adjusting on the proxy variable (T) instead of the unobservable/unmeasurable confounder variable (Z). For the binary case, it corresponds to the difference $StatDisp_T(Y,A) - StatDisp_Z(Y,A)$. For the linear case, it corresponds to the difference between the partial regression coefficients $\beta_{ya.t} - \beta_{ya.z}$. The difference between the adjustment free estimation of $StatDisp(Y,A)$ (the regression coefficient β_{ya} in the linear case) and $StatDisp_T(Y,A)$ ($\beta_{ya.t}$) corresponds to the total of both confounder and measurement biases.

[13] Notice that, by the backdoor formula, $StatDisp_Z(Y,A)$ coincides with $ACE(Y,A)$.



((a)) Confounding and colliding bias.



((b)) Confounding, colliding, and measurement bias

Figure 11.22: Confounding, colliding and measurement bias.

Selection and measurement biases. Figure 11.22(b) shows the simplest case where measurement and selection biases occur simultaneously. Adjusting on both the proxy (T) and the collider (W) variables ($StatDisp_{TW}(Y, A)$ and $\beta_{ya.tw}$) leads to both types of biases occurring simultaneously. Substracting $StatDisp_Z(Y, A)$ (respectively β_{ya}) from $StatDisp_{TW}(Y, A)$ (respectively $\beta_{ya.tw}$) coincides with the sum of selection and measurement biases in the binary and linear cases respectively.

Confounding, selection, and measurement biases. In the same simple case of Figure 11.22(b), the difference between adjusting on variables T and W on one hand and adjusting on Z on the other hand ($StatDisp_{TW}(Y, A) - StatDisp_Z(Y, A)$ in the binary case and $\beta_{ya.tw} - \beta_{ya.z}$ in the linear case) encompasses the three types of bias.

Confounding and interaction biases. In presence of interaction between two sensitive variables, confounding bias can be decomposed into interaction free portion and an interaction term. Figure 11.23(a) shows a simple confounding structure between A and Y and a second sensitive variable B which is interacting with the effect of A on Y. In the binary case, the confounding bias $ConfBias(Y, A)$ (Definition 11.2.1) can be decomposed as follows:

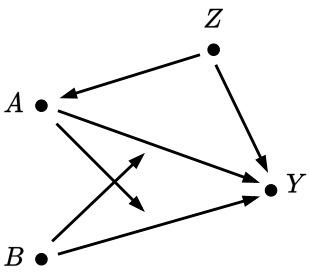
PROPOSITION 11.16.

$$\begin{aligned} ConfBias(Y, A) &= StatDisp(Y, A) - StatDisp_Z(Y, A) \\ &= SD_{\cancel{Int}}(Y, A) - SD_{\cancel{Int}_Z}(Y, A) \\ &\quad + P(b_1)(Interaction(A, B) - Interaction_Z(A, B)) \qquad (11.60) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (11.61) \end{aligned}$$

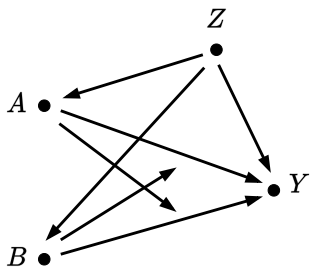
where

$$SD_{\cancel{Int}_Z}(Y, A) = \sum_Z (P(y_1|a_1, b_0, z) - P(y_1|a_0, b_0, z))P(z)$$

$$\begin{aligned} Interaction_Z(A, B) = \sum_Z \big(&P(y_1|a_1, b_1, z) - P(y_1|a_0, b_1, z) \\ &- P(y_1|a_1, b_0, z) + P(y_1|a_0, b_0, z)\big)P(z) \end{aligned}$$



((a)) Interaction and confounding bias



((b)) Interaction and confounding bias

Figure 11.23: Interaction and Confounding bias.

In the same example of Figure 11.23(a), the confounding bias in case of intersectionality (two interacting sensitive variables) can be decomposed as follows:

PROPOSITION 11.17.

$$\begin{aligned} ConfBias(Y,A,B) &= StatDisp(Y,A,B) - StatDisp_Z(Y,A,B) \\ &= SD_{\cancel{Int}}(Y,A) - SD_{\cancel{Int}_Z}(Y,A) \\ &\quad + Interaction(A,B) - Interaction_Z(A,B) \qquad (11.62) \\ & \qquad (11.63) \end{aligned}$$

In the slightly different structure where Z is also a confounder between B and Y (Figure 11.23(b)), the term $SD_{\cancel{Int}}(Y,B) - SD_{\cancel{Int}_Z}(Y,B)$ needs to be added to the $ConfBias(Y,A,B)$ expression above.

12

Gender and Sex Bias in COVID-19 Data

12.1 Introduction

Sex and gender disparity was noticed in many cases of Coronavirus disease 2019 (COVID-19). In this chapter we follow the definition proposed by [241] distinguishing between *sex* as a set of biological attributes, and *gender* as a social-psychological category. As it is later demonstrated, both might have an impact on COVID-19 mortality rates. We also note that in this study we consider binary values for gender, although an in depth analysis of gender roles would potentially yield a more complex picture. The disease is reported to be deadlier for infected men than women with a 2.8% fatality rate in Chinese men versus 1.7% in women [242], while sex-disaggregated data for COVID-19 in several European countries shows a similar number of cases between sexes, but more severe outcomes in aged men [242].

Biological differences in the immune system in men and women may affect the person's ability to fight COVID-19. It may be argued that men are more vulnerable to COVID-19 in relation to women because of a distinctive lifestyle, smoking, drinking, working hours, sex hormones, hypertension, and other circumstances [243]. Research suggests sex-based differences in ACE2 and TMPRSS2 enzymes and the link between circulating ACE2 and COVID-19 [242] is not clear. Additionally, sex and gender may intersect with age and race, to further increase the risk of severe COVID-19 outcomes in men. In *PLoS pathogens* and *CMAJ* journals it is also discussed how other socio-economic factors also increase the risk of COVID-19 [244, 245]. Systemic health and social inequities have disproportionately exposed low-income communities, racial and ethnic minorities to higher risk of COVID-19 infection and death. Additionally, uneven testing strategies across the world, and the quality of epidemiological big data, limit the accuracy of estimated distribution of COVID-19 patients according to [246].

The contributing factors can broadly be categorized into physical, sex-related attributes, lifestyle gender related attributes, and cultural, gender role related variables. The contribution, consistency of the effect and causal role of each of those groups of variables is very different and must be taken into consideration when building machine learning models, performing data analysis or making data-informed decisions. While physical, sex-based factors can be viewed as relatively constant predictors, the gender lifestyle attributes are fluid and vary from individual to individual. Furthermore, the cultural, gender roles based variables are intrinsically contextual and culture specific. Failing to account for those individual and cultural differences hinges both the accuracy of the predictions and put the group of individuals under a threat of disparate impact of such predictions. However, the complex structure of the various factors that influence the disease may not be evident from the accessible health databases. Observational datasets coming from public information access systems can be fragmented, coming from diverse sources and may not necessarily include all the attributes relevant for the analysis. This chapter showcases potential risks of biased or incomplete data and how causality can be put into practice as part of a risk management strategy to avoid discriminating systems. In this chapter we focus on analyzing the difference in causal and fairness impacts of different categories of variables linking sex or gender and COVID-19 vulnerability. We demonstrate how omitting causal, research-based knowledge from the model of sex and COVID-19 relationships can further propagate more intricate forms of bias in computational models and lead to discriminatory and harmful pandemic policies and decision making.

As a result, we bring light into: 1) a potential set of hypotheses within our COVID-19 case study to further verify its causal link, and 2) the unintended consequences that can derive from a lack of an adequate toolbox to support fair and accurate decisions.

12.2 Related Work: Identifying causal explaining factors on sex/-gender and COVID-19 relationship from epidemiological and clinical studies

We review findings based on big data on gender and COVID-19 from two angles. First, we analyze a body of papers placing gender and sex as a risk factor towards COVID-19, focusing on explaining the reasons behind disparity. Second, we categorize the explaining variables into mediators and confounders and discuss possible fairness implications of the former results that could lead to discrimination decisions, with the aim of guiding the causal design of the underlying model.

The amount of literature providing evidence on links between sex/gender and COVID-19 vulnerability is significant [247]. Table 12.1 shows articles finding men to be more vulnerable to COVID-19 in comparison to women. The explanations for this association are as well diverse. One of the possible factors is sex impact on vaccine acceptance, responses, and outcomes [242]. Women are often less likely to accept vaccines but once vaccinated, develop higher antibody responses [248]. For example, after vaccination against influenza, yellow fever, rubella, mumps,

measles, small pox, hepatitis A and B and dengue viruses, protective antibody responses are twice as high in adult females compared with males. However they report more adverse reactions to vaccines than males [242]. Moreover, biological differences in the immune systems of men and women exist, and they may affect the capacity to fight COVID-19 infection. Men appear to be at a greater risk with COVID-19 compared to women, whose higher immunologic response is probably associated with decreased mortality [249]. Furthermore, certain differences in cardiac manifestations in COVID-19 must be considered as a core component [250]. From the observational studies perspective, men appear to be at a greater risk. Sex is surely not the only risk factor in a disease that, according to [243], is challenging to diagnose and theorize, and whose effects also depend on vulnerabilities related to diabetes, obesity, hypertension, heart disease, chronic kidney disease, and chronic pulmonary disease according to [244]. Many authors suggest that women naturally produce more types of interferon, which limits the abnormal immune response in the form of serious cases of COVID-19. Moreover, women also produce more T lymphocytes which kill infected cells; and the "female" hormone estradiol would also offer greater protection against infection. On the contrary, studies indicate testosterone would limit the immune response in men, which may explain the observed sex-bias [251, 252].

Immunitary response duration was studied at the Pasteur Institut[1] and CHU of Strasbourg on 308 healthcare personnel that developed a light form of COVID-19 [253]. They show significantly steeper, i.e., faster decline in antibodies (anti-S and NAbs) in males than in females independently of age and BMI, hinting to a lower duration of protection after SARS-CoV-2 infection or vaccination. As more protective antibodies are formed in women, they last longer and so, women are better protected.

The relevance of gender norms, roles, and relations that influence women and men differential vulnerability to infection, exposure to pathogens, treatment received, as well as how these may differ among different groups of women and men is outlined in [254].

When comparing the COVID-19 case fatality rate (CFR) between China and Italy, the authors in [255] infer how methods from causal inference –in particular, mediation analysis–, can be used to resolve apparent statistical paradoxes and other various causal questions from data regarding the current pandemic. Many research studies [256] revealed that systemic health and social inequities have disproportionately increased the risk of COVID-19 infection and death among low-income communities and racial and ethnic minorities. The outcomes in [257] provide insights on the clinical aspects of the disease, on patients' infection and mortality risks, on the dynamics of the pandemic, and on the levels that policymakers and healthcare providers can use to alleviate its toll. In the gender and social norms side, a recent study conducted in Spain (one of the hardest hit countries in Europe) reported that women had more responsible attitude towards the COVID-19 pandemic than men [258], and another in the US showed that women take more precautions, wear more masks and cover more coughs than men[2]. Gender roles are considered as those influencing women's and men's different vulnerability to infection and exposure to pathogens, as reported in [254]. The impact of gender-specific lifestyle, health

[1] https://www.pasteur.fr/fr/espace-presse/documents-presse/COVID-19-duree-reponse-immunitaire-neutralisante-plus-longue-femmes-que-hommes

[2] https://hbswk.hbs.edu/item/the-covid-gender-gap-why-fewer-women-are-dying

behavior, psychological stress, and socioeconomic conditions on COVID-19 is further studied in [242].

Study	Tested Hypothesis	Men are more vulnerable	Health Conditions	Age Correlation	Drinking / Smoking
Impact of sex and gender on COVID-19 outcomes in Europe [242]	COVID-19 is deadlier for infected men than women	✓(NSD: M)	C, H, DM, CD, CRD, CLD	✓	✓(D, S)
Coronavirus: why men are more vulnerable to COVID-19 than women? [259]	There are higher morbidity and mortality rates in males than females	✓(NSD: M)	O, DM, H	✓	
Biological sex impacts COVID-19 outcomes [244]	Mechanistic differences including the expression and activity of ACE2 enzyme result in antiviral immunity, cases, hospitalizations and deaths differences.	✓(NSD: M)	CPD, CKD, II, HD, O	✓	
COVID-19: the gendered impacts of the outbreaks [254]	Men are more likely to remain hospitalized an die and less likely to be discharged from the hostpital than women.	✓(NSD: M)	H	✓	✓(S)
Racial and gender based differences in COVID-19 [246]	Ethnic differences influence susceptibility and mortality	✓(NSD: M)	HD, O, CLD, C, H, DM, CD	✓	✓(D, S)
Sex Differences in Mortality From COVID-19 Pandemic: Are Men Vulnerable and Women Protected? [250]	Male sex plays a role in increased mortality rates	✓(NSD: M)	H, DM, CD, CRD, CLD	✓	
The influence of sex and gender domains on COVID-19 cases and mortality [245]	Gender Inequality Index is positively associated with male:female cases ratio	✓(SSD: M)	19		
Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission [251]	Male sex is a risk factor for death and ITU admission but not for infections.	✓(SSD: M)	H, II, C	✓	✓(S)

Table 12.1: Summary of claims involving statements regarding men being more affected by the COVID-19 compared to women. X indicates correlation of that variable with the COVID-19. M: men are more affected, F: women are more affected by COVID-19, SSD: Statistically Significant Difference, NSD: Non Statistically-significant difference. Factors: S: Smoking, D: Drinking, C: Cancer, H: Hypertension, DM: Diabetes mellitus, CD: Cardiovascular diseases, CRD: Chronic respiratory disease, CLD-chronic lung disease, HD: Heart disease, O: Obesity, II: Inflammatory immune responses, CHK: Chronic kidney disease, CPD: Chronic pulmonary disease. Even though most articles claim men are more affected by COVID-19 than women and die more, none of them shows statistical significance nor has enough data to provide causal links beyond correlational studies.

According to most of the literature observed, the sex bias observed in COVID-19 as stated by [251], is a worldwide phenomenon suggested by observational and clinical research. However, the explaining factors listed in the discussed studies are diverse and non-uniform in their sensitivity to individual or cultural contexts.

12.3 Gender-related lifestyle habits and COVID-19 vulnerability

To help understand the association between gender and COVID-19, we conducted a more focused data analysis based on publicly available data to investigate the possible impact of sex and gender on the COVID-19 epidemic[3]. Even if ecological analysis[4] is considered the lowest form of epidemiological evidence, and potentially involves confounding variables, we are aware that it may not be a more accurate assessment than the individual level studies being surveyed in this article. Nonetheless, in this section we use this kind of analysis in order to elucidate plausible risk factors and unaccounted variables potentially explaining the disproportionate results.

We constructed a database that aggregates confirmed cases statistics, COVID-19 deaths, ICU admissions and smoking data per gender for 61 countries spanning 5 continents. The data sources are briefly described below.

- The *Global Health 50/50*[5] project housed at University College of London, which is created by a live tracker that aggregates data on COVID-19 cases and mortality from published government reports. At the time of our analysis on April 05, 2021, sex-disaggregated data for 183 countries including confirmed cases, confirmed deaths, etc. was represented in the live tracker.

- We also used a public dataset maintained by *Our World in Data*[6], which also contains additional information such as smoking, population, and daily COVID-19 cases.

By aggregating data from these two sources, and including only countries for which confirmed cases, deaths and smoking information is available. It is worth noting that, in this analysis, due to missing data for some countries, and taking into account the low granularity of the data, our choice was to focus only on the countries where all data columns were complete. We were able to analyze complete data from 89 countries. [7]

We then looked at the *male-to-female* (male/female) ratio of confirmed cases, ρ_{cases}, and compared it to the *male-to-female* ratio of deaths, ρ_{deaths}, for each country. We particularly

[3] Data analysis notebook in R available for reproducibility online: https://rpubs.com/wafaeB/684506
[4] Studies where individual features and outcomes are aggregated at a group level and then analyzed.
[5] Global Health 50/50 project website https://globalhealth5050.org/
[6] Our World in Data portal ourworldindata.org
[7] The total aggregated multi-source data contained the following countries: Albania, Tunisia, Mozambique, Montenegro, Cyprus, Bosnia and Herzegovina, Spain, Turkey, Romania, Netherlands, Argentina, France, Portugal, Switzerland, Iceland, Kyrgyzstan, Sweden, Poland, Latvia, Eswatini, Jamaica, New Zealand, Croatia, Cambodia, Armenia, Ukraine, Slovakia, Belgium, South Africa, South Korea, Canada, Hungary, Vietnam, Slovenia, Mongolia, Lithuania, Estonia, Bahamas, Qatar, Thailand, Malawi, Burkina Faso, Bangladesh, India, Pakistan, Nepal, Nigeria, Yemen, Congo, Oman, Kenya, Panama, Costa Rica, Singapore, Dominican Republic, Liberia, Myanmar, Morocco, Bahrain, Haiti, Mexico, China, Greece, Philippines, Maldives, Paraguay, Zimbabwe, Colombia, Denmark, Italy, Barbados, Sri Lanka, Ecuador, Malta, Iran, Rwanda, Finland, Brazil, Indonesia, Israel, Austria, Chile, Norway, Luxembourg, Germany, Australia, Lebabnon, Uganda.

classified countries on 4 groups based on these two parameters as follows (Table 12.2):

- Group 1 includes the countries in which $\rho_{cases} > 1$ and $\rho_{deaths} < 1$. In our analysis, only two countries belong to Group 1.
- Group 2, which contains 32 countries, represents countries in which $\rho_{cases} < 1$ and $\rho_{deaths} > 1$.
- Group 3 contains 7 countries in which $\rho_{cases} < 1$ and $\rho_{deaths} < 1$.
- Group 4 includes 49 countries in which $\rho_{cases} > 1$ and $\rho_{deaths} > 1$.

More Deaths: *More Cases:*	Females	Males
Females	Group 3	Group 2
Males	Group 1	Group 4

Table 12.2: Summary of analyzed male-to-female cases ratio and male-to-female deaths ratio.

Among the analyzed countries in our study, only Lebanon and Uganda belong to Group 1. Our analysis revealed that while there are more confirmed cases among men compared to women, i.e. $\rho_{cases} = 1.45$ in Lebanon and $\rho_{cases} = 2.18$ in Uganda, the male-to-female ratio of deaths is still smaller, i.e. $\rho_{deaths} = 0.44$ and $\rho_{deaths} = 0.86$ in Lebanon and Uganda, respectively. Therefore, more deaths were reported among women. Thus, this case seems to be contradictory to global data which indicates that men are more likely to get severely affected by COVID-19, and die more from the disease than women. One of the possible reasons is women's representation in certain sectors strongly hit by the pandemic, such as the garment and textile sector, in some Asian and African countries. This can translate into two potential explanations motivating more deaths in women: 1) they become unemployed and without access to healthcare to deal with the disease, or 2) they become more vulnerable and most affected by cotton industry-related respiratory diseases related with the lack of safety equipment in unhygienic, unsafe environments with hazardous work conditions, as reported in [260, 261]. However, more research needs to be done in order to provide more insights on the vulnerability of women to COVID-19 in Vietnam.

In Group 2, which contains 32 countries, women were more contaminated by COVID-19 than men. However, the number of deaths among male was higher. Data for this group also shows that this might be related to the much higher smoking rate in these countries. As shown for example in Fig. 12.1, a very high male-to-female smoking ratios are observed in most of the countries in this Group. Particularly, the highest smoking rates are observed in Tunisia, Albania and Mozambique, which also have the highest smoking ratios. Note that, in Figures 12.1, 12.2, 12.3 we applied log scaling to the calculated ratios in order to plot them on a comparable scale. That is, a positive male-to-female smoking or death log-scaled ratio indicates a higher number of smoking or death among men, while a negative male-to-female smoking or death log-scaled ratio indicates a higher number of smoking or death among women. While smoking might be one of the reasons that increases the risk of hospitalisation and death by COVID-19, as it is the case for most respiratory diseases, more data is needed in order to provide evidence on this hypothesis, such as age, number of tests by gender, etc.

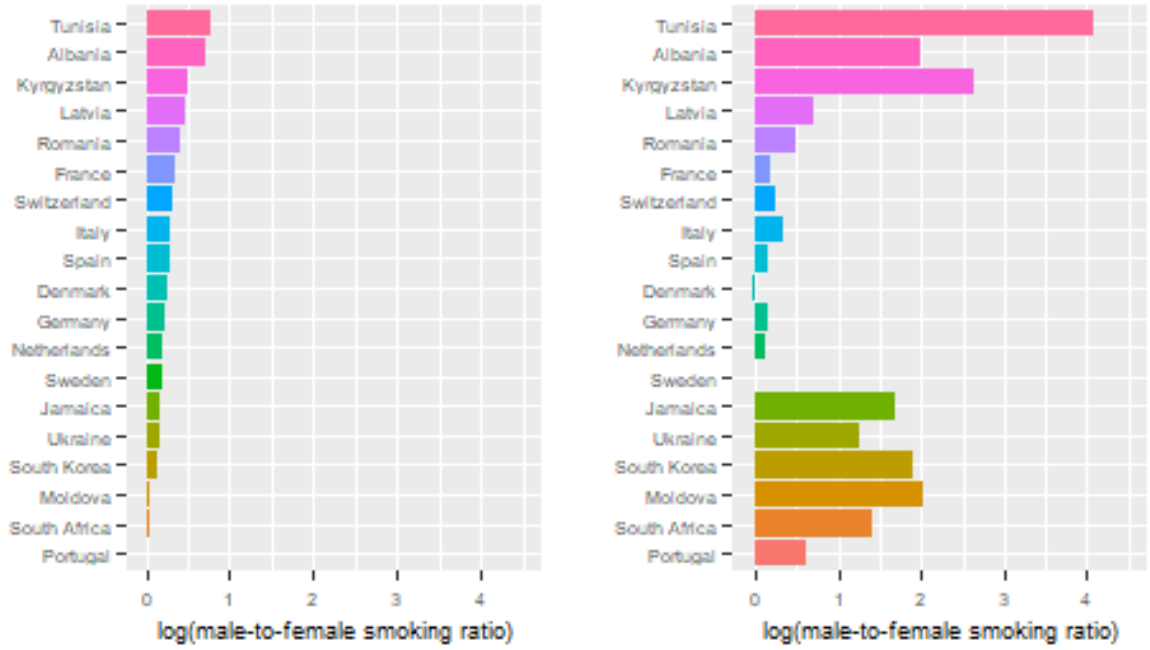


Figure 12.1: Log-scaled *Male-to-Female* Deaths ratio (Left) vs Log-scaled *Male-to-Female* smoking -female-to-male- ratio smoking (Right) for Group 2 (more cases in women but more deaths for men). This group is composed by 32 countries and shows that one possible explanatory variable is the factor *smoking*, since men are shown to smoke more in these countries.

Driven by the observations we made in the previous group of countries, we were also interested in investigating the association between deaths ratios and smoking ratios for Group 3 and 4. Fig. 12.2 and Fig. 12.3 report the *male-to-female* deaths ratio vs the *male-to-female* smoking ratio for Group 3 and Group 4, respectively. Group 3 represents 7 countries in which both confirmed and fatality rates are higher for women compared to men, while Group 4 represents 49 countries in which both confirmed cases and deaths are higher for men. Figures 12.2 and 12.3 also show a possible association between smoking and deaths. While the average *male-to-female* log-scaled smoking ratios is 1.5 across countries in Group 3, its value is higher and is up to 1.9 in Group 4, in which the *male-to-female* deaths ratios are also higher. It is also possible that countries in Group 3 are more likely to apply fairer testing strategies compared to the countries in Group 4, that have the highest *male-to-female* death ratios.

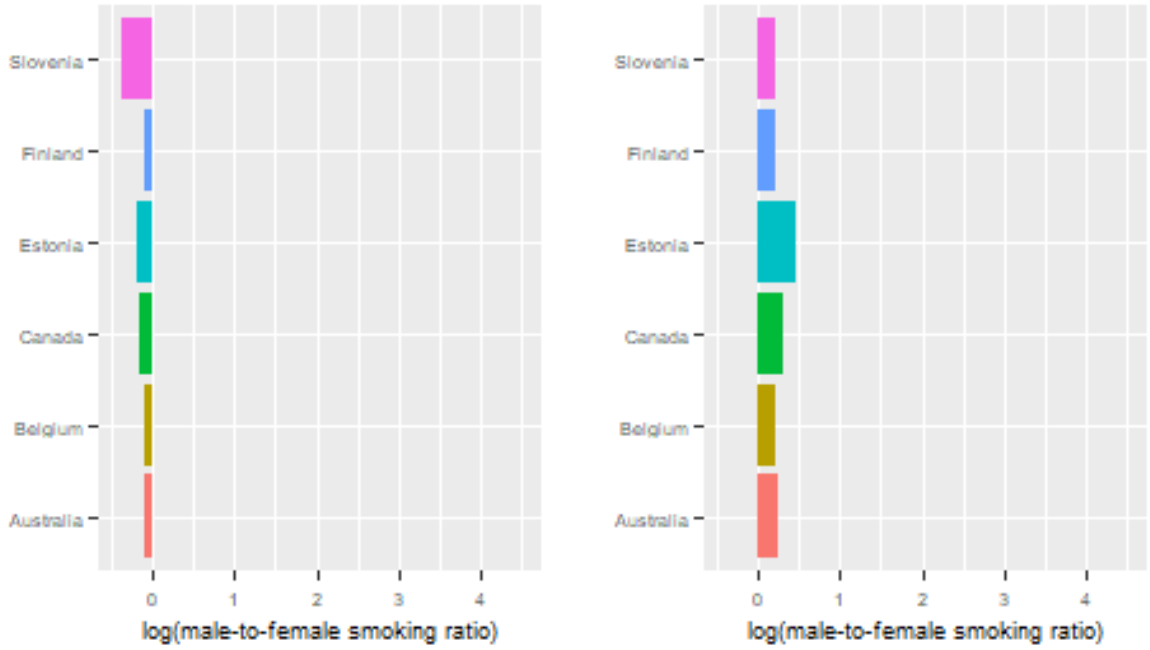


Figure 12.2: Log-scaled *Male-to-Female* Deaths ratio (Left) vs Log-scaled *Male-to-Female* Smoking ratio (Right) for Group 3 (7 countries, in which both cases and death ratios are higher for women, i.e., the opposite of most articles claims). In these countries, women smoke almost equally as men, and thus, smoking does not seem to clearly be an explanatory variable: women die as much or more than men.

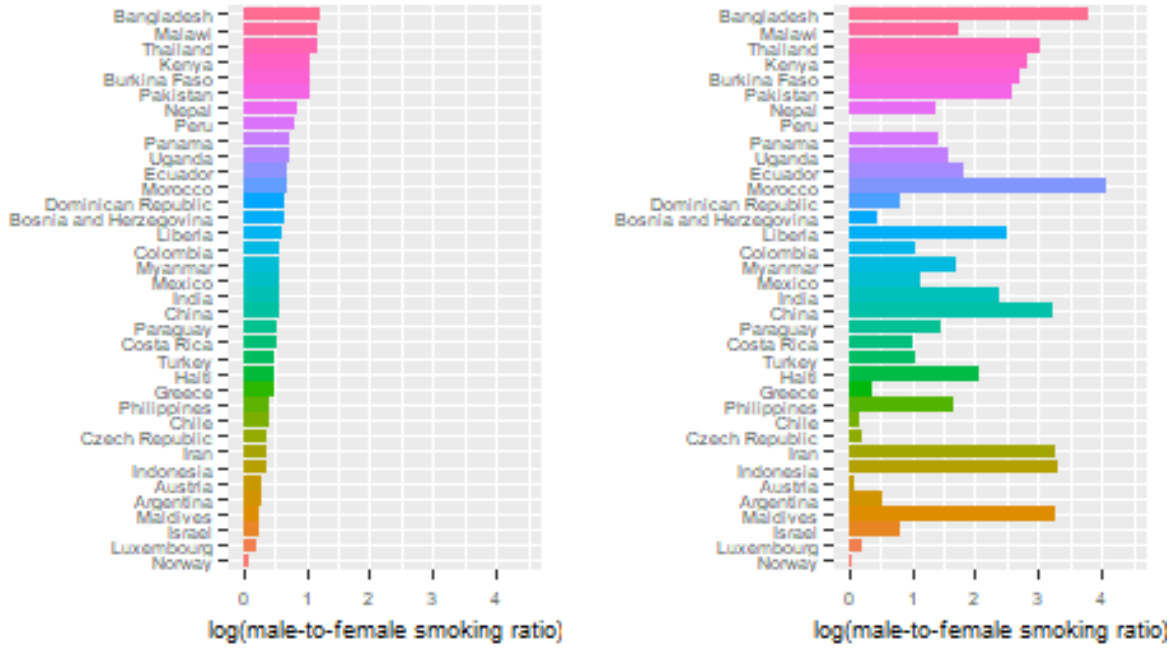


Figure 12.3: Log-scaled *Male-to-Female* Deaths ratio (Left) vs Log-scaled *Male-to-Female* Smoking ratio (Right) for Group 4 (49 countries where both cases and deaths are higher for men). This plot may reveal different testing strategies, as men are always more impacted.

While our analysis suggests a possible association between smoking and a higher number of COVID-19 deaths, as most countries having a high *male-to-female* deaths ratios, have a high *male-to-female* smoking ratios as well, there is no firm conclusion that can be drawn regarding

the relationship between smoking, sex and COVID-19. In addition, countries considered different criteria during the pandemic for reporting COVID-19 deaths and this could make understanding the impact of sex on COVID-19 ambiguous.

The differential findings and disparities observed across the four groups of countries in our analysis emphasize the need to understand why COVID-19 impacts some groups more than others. This might reflect other related factors and issues that need to be addressed, such as incomplete data and decision making biases. In the next sections we attempt to summarize the potential explaining variables found in the literature review and our analysis to structure it according to the ascribed role in the causal sex-COVID-19 relation framework.

12.4 Confounders and mediators between sex and COVID-19 vulnerability

In this section we summarize the variables linking sex or gender and COVID-19 vulnerability, and categorize them in mediators and confounders (Table 12.3). The mediators are further divided into constant (sex-related) and varying (gender-related) from individual to individual. It has to be noted, that this classification is dependent on the goal of the predictor, or a question that we are trying to answer. Which of the variables are considered confounders, as well as which mediators can be viewed as explaining variables (fair) or redlining (unfair) is context specific. For example, when predicting the probability of mortality from COVID-19 and allocating resources in the ICU (intensive care unit), the variable responsible for gender-related health-consciousness is a confounder. Namely, it does not directly influence the development of a disease in the hospital, but creates a spurious correlation in the epidemiological data. On the contrary, if the predictor was trying to answer the question which group should be more targeted by health-related social advertising (urge to wash hands or wear masks), the same variable could be used as an explaining mediator. Thus the men could be targeted more, proportional to a measured effect of the health-consciousness gender-related bias.

In Table 12.3, we consider COVID-19 severity and mortality risk as a prediction question, and healthcare resources allocation in the hospital as a decision based on the perceived level of severity.

Variable and Source	Class	Group	Comments
Hormones [244, 249, 251, 252, 262]	Mediator	sex-related Bio Var	Male hormone testosterone is associated with increased vulnerability, whereas female hormones are believed to play a protecting role.
Immune response [249, 253, 262]	Mediator	Sex-related Bio Var	More protective antibodies are formed in women and they last longer.
Smoking and drinking [242, 259]	Mediator	Gender-related Lifestyle Var	Higher smoking and drinking rates among men induce lung injuries that affect COVID-19 vulnerability.
Stress [242]	Mediator	Gender-related Lifestyle Var	Men often are more exposed to stress at work.
Hazardous industry [260, 261]	Mediator	Gender-related Lifestyle Var	It is worth noting that in some Asian countries women constitute a majority of garment and textile sector workers that are exposed to unsafe work conditions and are reported to be hit by the pandemics more than men.
Health behavior [242, 245, 258]	Confounder	Gender roles related Var	Women are more health-conscious and compliant with health recommendations
Exposure to pathogens [254]	Confounder	Gender roles related Var	In traditional societies women stay at home, and therefore are less exposed to the virus.

Table 12.3: Causal explaining variables between gender/sex and COVID-19 severity, classified into mediators and confounders. Mediators are the intermediate variables on the causal path from sensitive attribute to the outcome. A confounder is a variable with incoming arrows in the graph to both sensitive attribute and an outcome (a cause for both) and creates spurious non causal relationship between the two.

From observational analyses and tables in previous section we can observe a set of factors repeating as conditioning factors to explain the differences of sex and gender's impact on COVID-19 vulnerability. In this section we synthesize these factors to provide an overall aggregation of COVID-19-related claims most stated by the literature on the impact of different variables on COVID-19.

We are aware that other studies have considered other factors as important ones in the way COVID-19 infection translates into a severe case, for instance, the blood group type [263–265], vitamin D deficit [266, 267], or other genetic factors [268]. However, here we address only gender or sex related factors and their roles in predicting COVID-19 vulnerability.

Next sections will elaborate on the causal tools available to further study and corroborate such causal hypotheses and explanatory factors drawing on the body of analyzed literature.

12.5 Avoiding potential discriminating policies through a causal approach

In general, fair decision should not be based on any knowledge of the sensitive attribute such as gender, race, sexual orientation, etc. The case of medical diagnosis and treatment is an

exception, because certain diseases and conditions are specific to a particular sex, for example breast cancer which is almost exclusively characteristic for females. However, it is important to evaluate the exact extent of how much a physical component of being a female reduces the risk of mortality rather than gender related mediators or confounders. Current COVID-19 research shows that the underlying causes of vulnerability are diverse and unequal in the causal quality. As a consequence, this opens many pathways for the results to be distorted. Specifically, an already widely accepted discovery of women being more resilient to the disease can be affected by a spurious confounder or mediator which is not necessarily present in all women, and must be considered individually. Next, we illustrate unintended negative consequences for women if clinicians or governments base decisions on assumptions of greater resilience to the virus for females without adjusting for individual and cultural differences. We demonstrate the urge for more fine-grained causal analysis by performing mediation analysis on synthetic data generated following the epidemiological research informed causal model.

12.5.1 Data Generation and Model

To illustrate different causal paths between gender and COVID-19 severity we construct a causal model based on the discussed literature. We note that causal models can also be learned from data directly with causal discovery methods such as [64, 114, 269]. However, expert knowledge and previous research in the domain is important in informing what variables have to be included in the data. Furthermore, a recent study [223] shows that different causal discovery algorithms may not always agree on the resulting causal structure, therefore a combination of prior causal knowledge (for example, from experimental research) and statistical methods can help to achieve more robust results.

In Figure 12.4 we provide a Directed Acyclic Graph (DAG) to represent the causal structure of the data generating process. A DAG is a graphical representation of independence properties of joint probability distributions. It is constructed from the nodes that represent the variables and the edges that denote conditional probability relationships. In our case the joint probability of the variables in the DAG can be factorized as follows:

$$P(G,S,L,B,C) = P(G)P(S|G)P(L|S)P(B|S)P(C|L,B,G) \tag{12.1}$$

Where $P(G)$ is the probability of observing (different) Gender Roles (Equal or Traditional), $P(S|G)$ is the probability of entering the set of samples where $Sex = Female$ or $Sex = Male$ is observed within the infected patients given the value of Gender Roles, $P(L|S)$ is the probability of unhealthy lifestyle given Sex, $P(B|S)$ is the probability of biological factors (BioVar) serving as a protection against COVID-19 complications given Sex, and $P(C|B,L,G)$ is the probability of observing severe COVID-19 disease given BioVar, Lifestyle and Gender Roles. An important difference between a Causal Graph in Figure 12.4, and a Bayesian Network or Markov Chain is that parents of an edge are indicated based on assumed causal relationships [270] For example, despite the symmetric conditional independence relationship between symptoms and a disease (it is possible to predict symptoms given the disease or disease given the symptoms) the symptoms

cannot be denoted as a cause for a disease (in nature the disease, for example the infection, happens before the symptoms). As defined by [21], one of the most important properties of a causal DAG is that all nodes are independent of their non-descendants given their (immediate cause) parents. For example, given the model in Figure 12.4, COVID-19 Severity (S) becomes independent of Sex conditioned on all the intermediate parents such as BioVar (B), Lifestyle (L) and Gender (G) Roles. Formally:

$$P(C) \perp\!\!\!\perp P(S)|P(B, L, G) \tag{12.2}$$

The model indicates Sex as a "treatment variable" and COVID-19 as an "outcome" variable. That means we will estimate the effect of Sex on COVID-19 severity through various mediating and confounding paths. We must note that we include Sex, not Gender as a "treatment" variable, because Sex is usually included in the healthcare records. We distinguish aspects of cultural gender that are important for COVID-19 outcome as mediators and confounders and suggest that they should be explicitly taken into account when performing the analysis.

The model includes three groups of variables that serve as mediators or confounders between a sensitive attribute (sex) and the prediction (COVID-19 vulnerability) (Table 12.3). We consider *BioVar* as biological, sex-related attribute or a set of attributes such as differences in hormones, immune reactions and others, as one type of mediator variables that are constant (or almost constant) for biological *Sex*. The next group of mediators, *Lifestyle*, are related to sensitive attributes only through correlations between *Sex* and certain lifestyle choices such as smoking or drinking habits. Those attributes are gender-related. They can vary from individual to individual and cannot be automatically inferred from a *Sex* variable in the data.

Finally, the variable *Gender Roles*, account for spurious correlations between gender and COVID-19 severity and are considered confounders. This group is less intuitive to understand, because it is expressed with an incoming arrow from *Gender Roles* to the *Sex* variable, but *Sex*, as any sensitive variable, is considered to have temporal priority so it cannot be caused by other variables. However here we follow the Fairness Model by [199] and conceptualize *Gender Roles* variables not as causing *Sex* in the real world, but as causing the proportion of certain *Sex* values *in the sample* or a sampling bias. For example, traditionally, women are viewed as more careful and compliant with healthcare recommendations. This reduces the risk of getting the disease and the development of the disease under domestic treatment conditions. This results in less female cases among hospitalized individuals. However, it being more cautious has no effect on the further development of the disease when the patient is already in the hospital and is taken care of by medical staff. We also include a variable Y in our model to express a policy or treatment decision based on the predicted severity of the disease. We discuss the implications of the different combinations of paths causing the outcome and particular decision in Section 12.5.3. Note that we build our model only to illustrate different paths between *Sex* and the *Severity* of the disease, but not to predict the actual severity in the individual. Therefore we do not include other variables important for the COVID-19 outcome not related to *Sex*, such as non-gender-related health conditions. Some of the variables could be related to race or social status, and a more complex model is required to account for them. However, it goes beyond the scope of this article

and could be foreseen for future work.

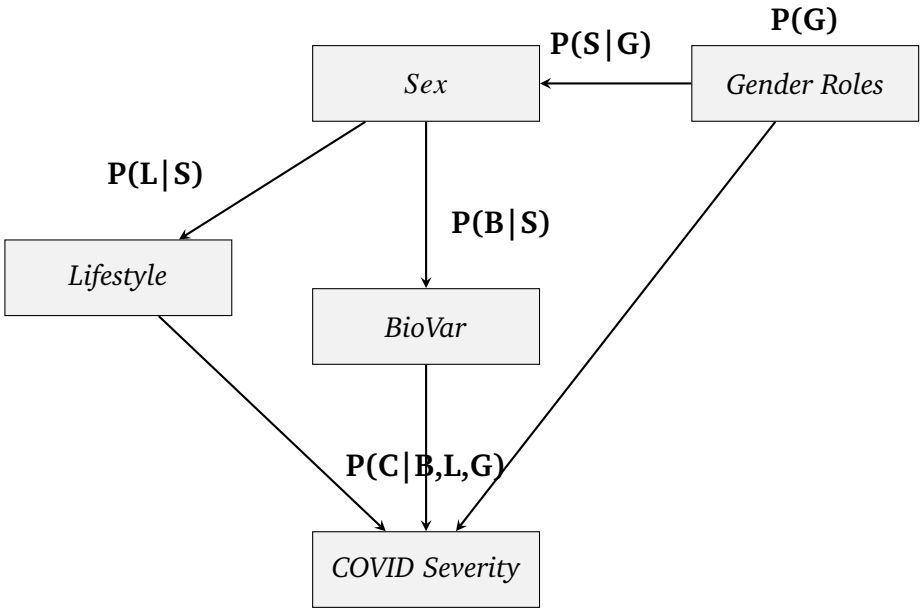


Figure 12.4: The model for the variables explaining the link between sex/gender and COVID-19 severity informed by the gender/sex related COVID-19 research.
P(G) - prior probability of observing a given Gender Role (Traditional or Equal). We make no assumptions and consider the probability to be equal for both values. However the prior probability is culture/country specific and can be informed from local sociological research. The prior and conditional probabilities for Bernoulli variables are explained in equations 12.3-12.7. The Lifestyle and BioVar variables are considered mediators (on the causal path from Sex to COVID-19 Severity). The Gender Roles variable is considered a confounder (spurious effect between Sex and COVID-19 Severity) and reads as "the probability of observing each value of sex among hospitalized individuals and the probability of severe development of the disease for each value of sex depending on the gender roles". The graph accounts for the fact that culture-based gender roles may be causing a particular sex to behave in a certain way that affects the way he/she gets exposed to COVID-19, and exaggerates the effect of sex on developing severe COVID-19 disease. We thus set as confounder the variable Gender Roles (it is a back-door path because it points both at the cause -Sex and the effect -COVID-Severity). Thanks to mediation analysis, which is a known method in causality, we can find the extent to which variable Sex has a direct and/or indirect effect on COVID-19-Severity only if we have real data. However, if the synthetic data was true, we could assert that the largest effect is through the mediator biological variables (Biovar: 0.521). However, a part of the effect is through the confounder variable GenderRoles (0.0541) and through the Lifestyle mediator (0.148) variables [8] . In this case the effect is not direct (because of the additional paths from Sex to Biovar and from Sex to Lifestyle) and part of it is not-causal (because of the confounding path through gender roles), opening the possibility for negative fairness implications, namely underestimating the COVID-19-Severity for women.

The relationship between variables encoded in a DAG provides the means of recognizing conditional independence and identifying the set of parameters needed for any given computation [271]. The model allows to identify the set of covariates for performing mediation analysis to evaluate the effect of Sex on COVID-19 Severity in the synthetic dataset (faithful to the model). The reason behind using synthetic data is twofold. First, the scenarios we are seeking to illustrate

[8]The path specific effects and the confounder effect add up to the total variation between severity for men and women, which is between -1 and 1 (See Section 12.5.3, where we illustrate with examples how, under this scenario, resource allocation without performing causal analysis would have negative fairness implications).

are related to the impact on individuals and require individual level data which is not freely available. Second, the purpose of this analysis is purely illustrative. Therefore the use of synthetic data and metrics derived from it help to convey the message without the danger of implying the usage of the results directly for clinical applications. The derivation of the real metrics for a particular use case has to come from quality individual level data that is representative of the local population and gender related cultural factors.

We generate binary data respecting the relations described in the model we built based on our review of gender/sex related COVID-19 literature and the groups of mediating/confounding variables we distinguish in Table 12.3. Figure 12.4 illustrates the dependencies between the variables expressed as prior and conditional probabilities. Note that for the sake of simplicity of illustration the possible variables from each group are expressed as one combined group variable. The prior and conditional probabilities ($P(G)$, $P(S|G)$, $P(L|S)$, $P(B|S)$ and $P(C|B,L,G)$) we assign to the variables are not based on estimations from the data, but we respect the causal directions described in the literature. For example, we set the probability of the protective value of biological variables (BioVar) to be almost coinciding with the female sex (0.99%) and non-protective value with male sex.

Similarly, the probability of healthy Lifestyle is higher for females, but it is less deterministic than the biological factors [242, 259], Gender Roles give rise to lower probability to observe females in the data (women get hospitalized less, perhaps because they take better precautions in daily life), as well as increased probability of mild rather than severe COVID-19 disease (Severity variable). Here we assume biological variables to have the largest overall effect on COVID-19 severity, lifestyle choices being the second, and gender role confounders as having the most moderate effect. The real proportions in the effect on COVID-19 severity can only be derived from a dataset including the relevant explaining variables for association between sex/gender and COVID-19 severity. We hope this article will encourage a causal analysis by proposing the model based on relevant research drawing attention to the relevance of causal knowledge for fair and explainable predictions.

The data is generated as follows. For simplicity all variables are set to be binary Bernoulli variables $\mathcal{B}$ with domain $k \in \{0,1\}$ and parameters $0 \leq p \leq 1$ and $q = 1-p$. The initial probability of the Gender Role variables being *Traditional* or *Equal* is set to be the same, namely 0.5 percent for each value. It can be adjusted based on our belief about a particular society where the data is collected.

$$GenderRoles \sim \mathcal{B}(0.5) \tag{12.3}$$

The Sex variable is set to be dependent on the Gender Roles variable. Namely, in the Traditional setting, women commute and are believed to be more health-conscious [242, 245, 254, 258], therefore, we observe overall smaller number of infected or severely ill female individuals. The conditional probabilities of Sex given Gender Roles reflect that hypothesis, but in absence of research on exact proportions, the numbers used are fictional. Under equal Gender Roles this effect is not observed, therefore the proportion of both sexes is equal.

$$Sex \sim (GenderRoles; p) = \begin{cases} Male : p_1 = 0.7, & \text{if Gender Roles} = Traditional, \\ Female : p_2 = 1 - p_1 & \\ Male : p_1 = 0.5, & \text{if Gender Roles} = Equal. \\ Female : p_2 = 1 - p_1 & \end{cases} \tag{12.4}$$

This means, that in case of Traditional Gender Roles the probability of getting infected (entering the sample) is much higher for men, whereas in the equal society the probability of getting sick is the same for both sexes.

The biological variables BioVar such as sex hormones or immune system specifics [244, 249, 249, 251–253, 262, 262] are treated as almost deterministically dependent on Sex. We acknowledge, that more research on the individual fluctuations of those paramaters would benefit the model.

$$BioVar \sim (Sex; p) = \begin{cases} Protective : p = 0.01 & \text{if Sex} = Male, \\ Protective : 1 - p & \text{if Sex} = Female. \end{cases} \tag{12.5}$$

Unhealthy lifestyle $value = 1$ (such as unhealthy lifestyle due to smoking, drinking, stress, etc.) are set to be more likely for men than for women [242, 259]. The exact proportion is not grounded in the literature and is for illustration purposes only.

$$Lifestyle \sim (Sex; p) = \begin{cases} Unhealthy : p = 0.7 & \text{if Sex} = Male, \\ Unhealthy : 1 - p & \text{if Sex} = Female. \end{cases} \tag{12.6}$$

This would mean that probability to observe a male leading unhealthy lifestyle is 70% compared to only 30% probability of encountering a female with the same unhealthy habits.

Finally, we define probability of COVID-19 Severity as a linear combination of the previously discussed variables. The proportions of the impact of each group of variables in the equation is motivated by the corresponding volume of the research supporting the hypothesis in the reviewed literature at the time when this study is performed.Note that linearity of the effect is only an assumption made for simplicity and does not imply the real interaction between different factors.

$$COVIDSeverity \sim (GenderRoles, BioVar, Lifestyle; p) =$$

$$\begin{cases} SevereCOVID - 19 : p = 0.2 \times GenderRoles + 0.5 \times BioVar + 0.3 \times Lifestyle \\ MildCOVID - 19 : 1 - p. \end{cases} \tag{12.7}$$

This means that the probability of severe COVID-19 disease is defined by the linear combination of the previously discussed variables.

However, the true functional form and exact proportions of the impact of each variable can be learned from the complete epidemiological data and is subject to future epidemiological research.

12.5.2 Mediation Analysis to analyse causal effects of sex on the severity of COVID-19

To determine the proportion of the effect each of the variables has on the severity of the disease we perform causal mediation analysis (Figure 12.6)[9]. Similarly to [255], where the proposed confounder is the age, we analyze the total effect and the effect of the mediating variables under the confounding variables of *Gender Roles*. We apply causal fairness notions such as Total Effect (TE 10.1), Natural Direct Effect (NDE 10.2) and Path Specific Effects (PSEs 10.4) through each mediator to determine the sex/gender bias in COVID-19 severity; namely, how much more likely is to observe a severe COVID-19 case for a man than for a woman.

To compute path-specific causal effects (PSEs) we use the imputation-based estimation of counterfactual outcomes implemented in R in the open-source Paths Library[10] [273] designed to trace causal paths from experimental and observational data. We use mediation analysis to estimate the proportion of the causal effect from Sex to COVID-19 Severity that is explained by one of the mediating variables. The imputation approach provides $K+1$ models that describe the expectations $\mathbb{E}[Y|\mathbf{X}, A], \mathbb{E}[Y|\mathbf{X}, A, M_1], ..., \mathbb{E}[Y|\mathbf{X}, A, M_k]$, where A is a sensitive attribute, $\mathbf{X}$ is a set of covariates, and $M_1, ..., M_k$ are mediators [19]. For more extensive explanations of the Causal Fairness Notions we refer the reader to the survey of Makhlouf et al. [19].

We also compute the Total Variation (another name is Statistical Disparity (SPD), Eq. 3.1). TV is a non causal fairness metric, and thus, it does not distinguish mediators from confounders. Note that in absence of confounders, TV and TE are equivalent. Hence, intuitively, the Confounding Effect [11] (CE) can be estimated by subtracting the Total Effect from the Total Variation: $CE = TV - TE$[12]. It is important to consider TV in our study as it corresponds to simple correlation between the Sex/Gender and the COVID-19 severity. In contrast, the remaining metrics are more fine-grained in considering a specific path between Sex/Gender and COVID-19 vulnerability.

All causal effects are obtained by subtracting the probability of severe COVID-19 being a man from the same probability while being a woman. Hence, a positive value indicates men are more likely to develop severe COVID-19 case, while a negative value indicates a COVID-19 severity bias for women. A value of zero means that the probability of the outcome is equal for men and women. A value equal to one or minus one would indicate extreme cases, where the probability of severe COVID-19 disease is equal to one hundred percent for one group and zero for the other. The Confidence Intervals (CI) for each value are calculated via bootstrapping methods included in *paths* library and indicate the significance of the effect [274].All effects except Natural Direct Effect indicate a severity bias for men (Table 12.4). The Natural Direct Effect is negative, close to zero, and the corresponding CI includes zero [13] indicating that the detected effect between Sex and COVID-19 severity is not significant.

[9]The code to generate the data and the analysis can be found in the repository https://github.com/RuSaBin/Covid_Gender

[10]https://github.com/cran/paths We direct the interested reader to the comprehensive survey on the libraries to perform mediation analysis in [272]

[11]Also known as Spurious Effect.

[12]This is not a formal definition, as the formula does not necessarily apply in non linear settings; however it is sufficient for illustrating the confounding effect in our data.

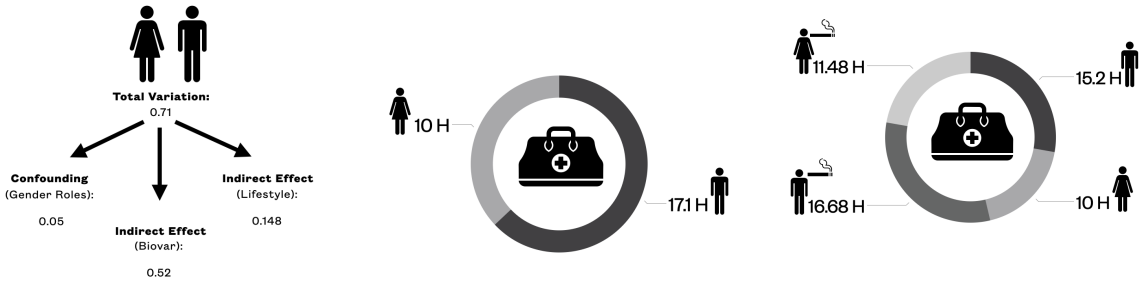
[13]which means that it is either small negative, small positive, or zero.

There is no Natural Direct Effect of Sex on COVID-19 because we assume that all the influence of sex/gender on COVID-19 severity is explained by the mediating variables (Table 12.3), even if some of them, such as BioVar (for example, female hormones) almost exactly correspond to sex. We make the decision to separate sex in general from specific sex related bio variables, to make it possible to account if needed, for individual fluctuations in those attributes and emphasize the more detailed explainability of the effect of sex on the disease severity. Following the interpretation of mediation analysis by [275] we discover that being a male increases the overall risk (Total Effect Equation 10.1) of severe case of COVID-19 by 65.4%. Note that this is different from the 70% estimated by the Total Variation (TV) before adjusting for confounding variables. Since the Direct Effect of Sex on COVID-19 severity is negligible (0.015, Equation 10.2) the causal effect that links Sex and COVID-19 severity is composed entirely of an indirect effect through BioVar of 52.1% and an indirect effect unhealthy through lifestyle of 14.8%.

We illustrate the contrast between estimating Total Variation or performing Mediation Analysis in Figure 12.5.

Let us say that we set a 10 hours minimum amount of hours of medical attention for hospitalized COVID-19 patients as a baseline. We want to allocate additional hours proportional to the risk of developing a severe COVID-19 outcome. In the case of computing Total Variation of effect of Sex on COVID-19 severity Figure 12.5(b)) we would allocate male patients 70% (TV = 0.7081 in Table 12.4) more of time, namely 17 hours. However, in the second case (Figure 12.5(c)), assuming that sex is almost a perfect proxy for biological variables (BioVar), we allocate male patients only 50.2% more of resources, namely 15.2 hours (given that women get 10 hours). Additional attention hours are allocated to smoking patients, regardless of sex, proportionally to the effect of smoking on severe COVID-19 disease: 11.48 hours for female smokers and 16.68 hours for male smokers (we add additional hours to the minimum based on path specific effect through lifestyle 0.148 in Table 12.4). Note that the synthetic data is generated assuming a conservative scenario, where the most significant part of the effect is due to biological variables which are closely correlated with sex (Table 12.3). If a larger part of the total effect was due to confounders such as gender roles relataed behaviour, the disparity between the Total Variation and Total Effect would further increase.

The amount of the effect caused by lifestyle or BioVar mediating variables, or the Gender Role confounding variable are not causally equivalent. In the following section we elaborate on the differences between BioVar or Lifestyle mediating variables, such as smoking or drinking habits and confounders responsible for spurious non causal effects, such as compliance with the healthcare recommendations (the variables and their belonging to the groups are listed in Table 12.3). We discuss the danger of failing to account for them in COVID-19 related policy making .



((a)) The decomposition of the Total Variation of severe COVID-19 risk for males and females.

((b)) The allocation of hours of medical care proportional to the estimate of Total Variation, namely if the minimum is 10h for women, men are allocated 70 percent more - 17 hours.

((c)) The allocation of hours of medical care proportional to the estimated path specific effects of Sex on COVID-19 Severity estimated by mediation analysis.

Figure 12.5: The Illustration of resource allocation according to different estimations of effect of Sex on COVID-19 severity via mediation analysis.

Variable	Estimated Effect	Standard Error	CI Lower 95%	CI Upper 95%
Natural Direct Effect: $Sex \rightarrow COVIDSeverity$	−0.015	0.035	−0.061	0.048
Path-Specific (Indirect) Effect : $Sex \rightarrow BioVar \rightarrow COVIDSeverity$	0.521	0.039	0.498	0.609
Path-Specific (Indirect) Effect: $Sex \rightarrow Lifestyle \rightarrow COVIDSeverity$	0.148	0.031	0.084	0.168
Total Effect: $Sex \rightarrow COVIDSeverity$	0.654	0.008	0.644	0.668
Total Variation: $Sex \rightarrow COVIDSeverity$	0.7081	-	-	-
Confounding Effect: $Sex \leftarrow GenderRoles \rightarrow COVIDSeverity$	0.0541	-	-	-

Table 12.4: Mediation Analysis of Causal Effects that illustrate the different paths of the influence of sex on COVID-19 severity. All effects except Direct Effect indicate a severity bias for men (positive values indicate severity bias for men, and negative values indicate severity bias for women). The Direct Effect is close to zero, because we assume through the causal graph used as prior model of the world that all the influence of sex/gender on COVID-19 severity is explained by the mediating variables (either BioVar or Lifestyle variables). The effect caused by BioVar mediating variable is higher than the effect caused by the Lifestyle mediating variable. The last two columns of the table indicate lower and upper bounds for confidence intervals for the estimated effect values.

12.5.3 Disparate impact of sex on COVID-19 treatment decisions

Decisions in real life based on biased data can create disparities in treatment or disparate impact resulting in disadvantage for protected groups.

Disparate treatment is a variation in decisions for individuals that depends on the values of a sensitive attribute. *Disparate impact* occurs when decision outcomes disproportionately benefit or hurt members of certain sensitive attribute value groups [276].

The adequate evaluation on fairness of decisions depends on the situation where the data analysis results will be applied. Here we would like to illustrate the evaluation of fairness in the

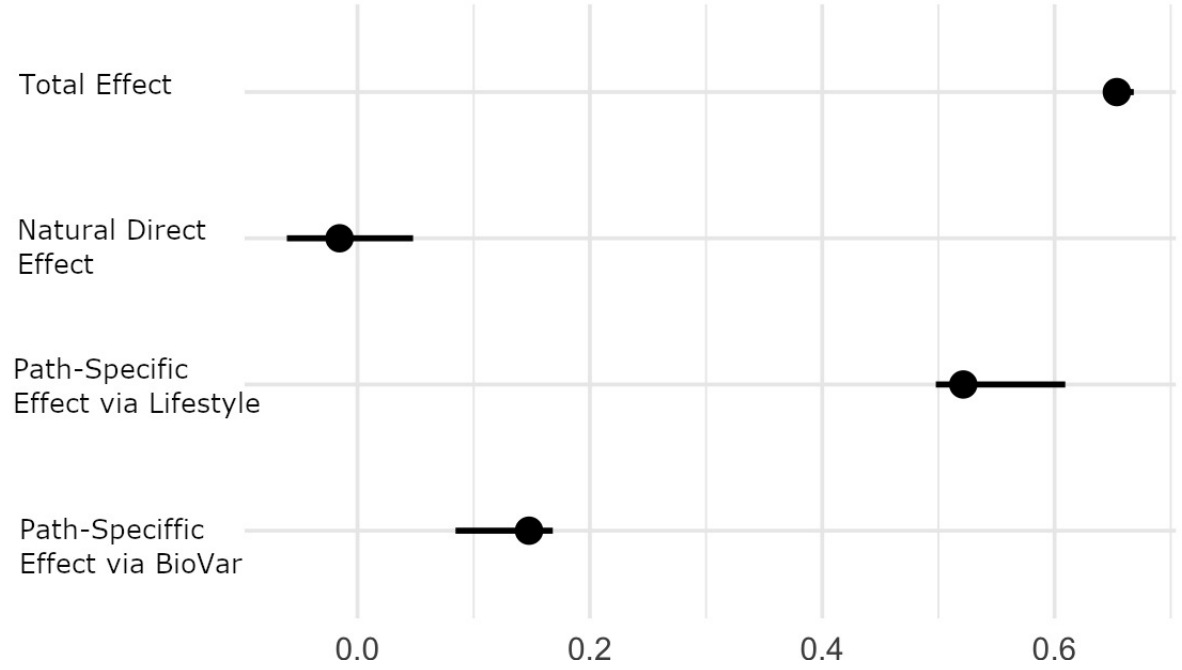


Figure 12.6: The graphical summary of fairness notions, Total Effect (TE), Natural Direct Effect (NDE), and Indirect Effects (NIE) on COVID-19 severity through biological (BioVar) and Lifestyle variables along with their confidence intervals. These metrics indicate the difference in the probability of a severe form of COVID-19 disease for men and women. Positive values indicate that Sex = Male is associated with higher probability of severe COVID-19 disease than Sex = Female. A negative value for NDE would mean the opposite, higher probability of severe COVID-19 disease for Sex = Female, however the small value is interpreted as not significant. A score of 0 means probabilities are equal. Confidence Intervals are calculated for regression based estimates of mediation analysis metrics (NDE, TE, NIE).

COVID-19 pandemic context, by modeling a situation where inference about women being less vulnerable to the virus is used for assigning a priority treatment to an individual (for example, a longer hospitalization, closer monitoring or priority access to vaccines).

We assume that higher vulnerability or risk of severe symptoms for men is inferred from observing more cases of hospitalized individuals in the electronic health records data. We will illustrate different implications on fairness when classifying individuals based on *Sex* in combination with three groups of variables: lifestyle mediators (Lifestyle), biological mediators (BioVar) and gender-roles (Gender Roles) confounders In reality other variables not related to sex, as well as other health condition indicating features, can influence individual vulnerability to the virus. Thus, a thorough causal analysis becomes even more relevant: decisions should be based on known causes of vulnerability rather than on a sensitive attribute.

Scenario 1: *Disparate impact due to Gender Roles confounding variables*

In this case the confounding variable *Gender Roles* indicates whether the member of a sensitive group follows traditional or equal gender behaviour models. Under the traditional setting we assume that women are more careful and compliant than men, which makes them less likely to get COVID-19, as well as more likely to improve their condition when sick at home However, once in a hospital, where the patient is taken care of by the medical professionals, the impact of being more careful diminishes. Furthermore, individuals that do not follow the traditional

gender-related behaviour might not fall into the same pattern. Failing to adjust the predictive model to this confounder bias, would predict women to be more protected from the severe COVID-19 disease forms than they really are. The Mediation Analysis on our synthetic data shows that men are expected to be more vulnerable than women 0.05 points more than they really are. Note that in reality the confounding bias can be much higher If the prediction is used to, for example, allocate limited resources in the ICU, women would be discriminated by being systematically denied priority treatment proportionate to the confounding effect.

Scenario 2: *Negative impact due to not accounting for Lifestyle mediator variables*

In this case the association between *Sex* and *COVID-19* cases is created by the mediating variable *Lifestyle* if it is not included in the data. *Lifestyle* choices such as smoking or drinking have a valid causal effect on severeness of the disease and thus, assigning a priority treatment to smoking individuals is adequate. However, if the prediction is based on *Sex* only, without observing individual patients' lifestyle habits, the women that are smokers would be wrongly classified as more resilient than they really are. As a consequence, they would be denied a part of necessary medical attention proportionate to the lifestyle Path-Specific Effect, i.e., 0.148 points higher estimated probability of severe outcome for men than for women (Table 12.1).

Scenario 3: *Using Biological mediators for sex-related COVID-19 severity prediction*

Considering the biological sex-specific variables such as hormones, adaptive immune systems and other variables, it is relatively safe to assume that their effect on the outcome of the disease can be predicted from the *Sex* variable. This allows for a unique situation, where using sensitive attributes is both allowed and necessary to ensure fair and accurate predictions. For example, insisting on identical treatment for men and women could result in disparate impact on health and mortality outcomes for men, proportionate to the *BioVar* Path-Specific Effect: it results in 0.521 percentage points higher probability of severe outcome for men (Table 12.4). Nevertheless, a careful causal path analysis is required to distinguish biological sex-related attributes from gender-related mediators or confounders that can bias the result and create unwanted discrimination at the individual or population level. In addition, individual fluctuations in biological markers can also supposedly affect constant sex-severity relationships through biological variables.

12.6 Discussion

The observed larger amount of hospitalized males in comparison with women, can be explained with several mediating and confounding variables. For instance, men lifestyle is different from women's, which can be a reason that men are more affected by COVID-19 infection. Men are more inclined towards drinking and smoking, which can evolve into lung infection which in turn, can formulate a larger chance of COVID-19 infection.

Social and cultural differences are additionally affecting the COVID-19 pandemic. In this line, another potential factor is the tendency of females to comply more with regulation, protecting themselves more and wearing masks more. Women are typically in charge of ensuring health for the whole family as part of their traditional reproductive work. Their greater compliance with COVID-19 recommendations is a reflection of long-established gender social roles, which has

also involved an increased burden for them during the pandemic [277]. However this behaviour does not impact the further outcome of the disease once in the hospital. Furthermore, this gender-related feature is not constant across individuals and populations.

Since the latest ML models such as deep networks do not correct against, but rather replicate existing biases of the researchers who train them, the data they are fed with, the circumstances of their testing, etc., we hope more effort is initially put into both performing representative data collection and causal data analyses. Likewise these checks need to be present when developing methods able to programmatically verify, flag, and reduce data and model biases. Stating the verified tests and/or including our recommended potential mediators and confounders will minimally set the state of affairs on the table, and therefore, highlight and make legal processes stand up for process automation. As a positive side effect, AI-based accountability will be more easily gained and traced. Fair data analysis is only the first step towards human-centric societies endowed with responsible AI systems that serve citizens and governments make use of data-informed policies more efficiently.

Part VI

Conclusions and Future Work

In conclusion, our work encompasses various dimensions of fairness, causality, and privacy in the context of machine learning and AI systems. Next, we provide our main findings and future directions corresponding to different aspects of ethical AI:

Bias correction

We have introduced the BaBE framework, which leverages domain-specific knowledge to perform data pre-processing, with the aim of achieving conditional statistical parity (CSP) and equal opportunity (EO), even when the explaining variable is latent. An essential feature of our approach is that it does not rely on the assumption of independence between the explaining variable and the sensitive attribute, which is particularly important in healthcare applications. One promising avenue for future research involves delving into the precision and accuracy of estimation, examining how these factors depend on the probability distribution matrices $\hat{\mathbb{P}}[E|Z,S]$ and their relationship to the matrices representing external knowledge, $\hat{\mathbb{P}}[Z|E,S]$. Another task is considering the multidimensional E variable. Collaborations with domain experts to formalize accurate measurements of these matrices will also be essential. We believe that our approach can serve as a bridge for interdisciplinary collaboration between domain experts and ML fairness practitioners, facilitating more equitable decision making in AI systems.

Understanding sources of algoritmic discrimination

We have attempted to distinguish sample size and underrepresentation biases and characterize how each of them influences algorithmic discrimination. In light of empirical analysis of benchmark datasets and using off-the-shelf classification algorithms, we made three important observations. First, discrimination metrics defined using AUC and ZOL (which consider the trade-off between precision and recall) are more resilient to sampling biases than discrimination defined using FPR and TPR (equal opportunity). Consequently, in the presence of limited size or imbalanced training data, it is recommended to use fairness metrics based on the trade-off between precision and recall (e.g. equalized odds [14]) to reliably estimate discrimination. Second, for regression problems, discrimination defined in terms of MSE is significantly affected by variance for extremely small or imbalanced training sets. Therefore, it is recommended to treat discrimination values with caution in such cases. Third, in case of tabular benchmark fairness datasets, contrary to the results in computer vision, collecting more samples of the extremely underrepresented group according to the population distribution will typically amplify discrimination rather than reducing it. This result suggests that for tabular data pre-processing methods which equalize the sensitive distributions are preferable to collecting more data for a minority group. We see a more in-depth comparison between underrepresentation bias in computer vision, natural language processing, and tabular data as an interesting direction.

Causality and privacy

We draw attention to the effect of data privatization on the learnability of causal graph from data. To allow the comparison between two distinct privacy notions, namely LDP and local d-privacy,

we introduced a unified privacy measure based on an attacking perspective. Our exploration has shed light on the advantages of local d-privacy for causal discovery. Future research directions include extending the analysis to continuous data and understanding the impact of sample size on output metrics. Finally, we aim to design a locally private mechanism tailored to causal discovery tasks. We see the adoption of the d-privacy framework as a promising direction for this task.

Causality and legal practice

In the context of employing statistical causality tools for AI fairness, we consolidate three main arguments: accurately measuring discrimination free from causal biases, obtaining deeper insights through mediation analysis, and aligning fairness analysis with court practices. In addition, we discuss the challenges of applying causality in practice and distinguish the availability of a causal graph as an important step of causal analysis. We see a broader analysis of statistical causality in relation to European AI legislation as an important direction for future research. A particularly important direction is to propose and test practical approaches to apply causal analysis of the decisions of ML algorithms, when the explainability of the decisions is limited.

Causal biases and accurately measuring fairness

We provide closed-form expressions of a specific class of biases, namely causal biases. By analyzing the magnitude of bias in terms of the model parameters, we could establish an intuitive interpretation of bias based on the causal graph structure underlying each type of bias. We aim to further explore cases where multiple types of bias coexist. Another extension is exploring the interaction bias in the context of mediation analysis.

Causality for explainability

We discussed several hypotheses for the higher male than female vulnerability to the COVID-19 virus. We proposed a framework for including mediator and confounder variables corresponding to the hypotheses identified in the literature in fair and explainable prediction models. We used a toy model to illustrate both conceptually and numerically the impact of failing to do so on the fairness of the disease severity predictions. We acknowledge that conclusive research explaining male vulnerability to COVID-19 virus is not yet available. Incorporating future research on hormonal, inflammatory, immunological and phenotypical dimensions in severe COVID-19 disease is necessary for building fair, explainable and accurate models. Another interesting avenue for future research is further exploring mediation analysis to separate gender and sex variables in healthcare-related ML models. We believe that distinguishing biological and cultural aspects of gender in a healthcare care context can help to achieve more precise predictions and better generalizability of a model.

In summary, our work contributes to ongoing efforts to improve fairness, explainability, and protect privacy in machine learning and AI systems. We propose insights and novel interdisciplinary approaches for treating fairness and privacy, while being aware of the big picture of ethical AI. We believe that a holistic approach to ethical AI represents an important direction of AI

development, and our findings can guide future research and practices toward better data-driven decision making.

References

[1] K. R. Varshney. *Trustworthy machine learning*. Chappaqua, NY (2021).

[2] A. Jobin, M. Ienca, and E. Vayena. *The global landscape of ai ethics guidelines*. Nature machine intelligence **1**(9), 389 (2019).

[3] L. F. Wightman. *Lsac national longitudinal bar passage study. lsac research report series.* (1998).

[4] T. Brennan, W. Dieterich, and B. Ehret. *Evaluating the predictive validity of the compas risk and needs assessment system*. Criminal Justice and behavior **36**(1), 21 (2009).

[5] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. *Dissecting racial bias in an algorithm used to manage the health of populations*. Science **366**(6464), 447 (2019).

[6] J. Buolamwini and T. Gebru. *Gender shades: Intersectional accuracy disparities in commercial gender classification*. In *Conference on fairness, accountability and transparency*, pp. 77–91 (PMLR, 2018).

[7] A. Lambrecht and C. Tucker. *Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads*. Management science **65**(7), 2966 (2019).

[8] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. *Language (technology) is power: A critical survey of bias" in nlp*. arXiv preprint arXiv:2005.14050 (2020).

[9] X. Ferrer, T. van Nuenen, J. M. Such, M. Coté, and N. Criado. *Bias and discrimination in ai: a cross-disciplinary perspective*. IEEE Technology and Society Magazine **40**(2), 72 (2021).

[10] D. Danks and A. J. London. *Algorithmic bias in autonomous systems.* In *Ijcai*, vol. 17, pp. 4691–4697 (2017).

[11] A. Nielsen. *Practical fairness* (O'Reilly Media, 2020).

[12] S. Ruggieri. *Using t-closeness anonymity to control for non-discrimination* p. 31 (2014).

[13] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. *Fairness through awareness* (2011). 1104.3913.

[14] M. Hardt, E. Price, and N. Srebro. *Equality of opportunity in supervised learning* (2016). 1610.02413.

[15] T. Calders and S. Verwer. *Three naive bayes approaches for discrimination-free classification*. Data Min. Knowl. Discov. **21**, 277 (2010).

[16] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. *Counterfactual fairness* (2018). 1703.06856.

[17] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. *Avoiding discrimination through causal reasoning* (2018). 1706.02744.

[18] S. Barocas and A. D. Selbst. *Big data's disparate impact*. California law review pp. 671–732 (2016).

[19] K. Makhlouf, S. Zhioua, and C. Palamidessi. *Survey on causal-based machine learning fairness notions*. arXiv preprint arXiv:2010.09553 (2020).

[20] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. *Sex bias in graduate admissions: Data from berkeley*. Science **187**(4175), 398 (1975).

[21] J. Pearl. *Causality* (Cambridge University Press, Cambridge, 2009). URL https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B.

[22] W. Pan, S. Cui, J. Bian, C. Zhang, and F. Wang. *Explaining algorithmic fairness through fairness-aware causal path decomposition*. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1287–1297 (2021).

[23] K. Makhlouf, S. Zhioua, and C. Palamidessi. *On the applicability of machine learning fairness notions*. ACM SIGKDD Explorations Newsletter **23**(1), 14 (2021).

[24] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. *The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making*. Communications of the ACM **64**(4), 136 (2021).

[25] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. *A comparative study of fairness-enhancing interventions in machine learning* (2018). 1802.04422.

[26] C. C. Porter. *De-identified data and third party data mining: The risk of re-identification of personal information*. Shidler JL Com. & Tech. **5**, 1 (2008).

[27] A. Narayanan and V. Shmatikov. *Robust de-anonymization of large sparse datasets*. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125 (IEEE, 2008).

[28] L. Sweeney. *k-anonymity: A model for protecting privacy*. International journal of uncertainty, fuzziness and knowledge-based systems **10**(05), 557 (2002).

[29] N. Li, T. Li, and S. Venkatasubramanian. *t-closeness: Privacy beyond k-anonymity and l-diversity*. In *2007 IEEE 23rd international conference on data engineering*, pp. 106–115 (IEEE, 2006).

[30] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *l-diversity: Privacy beyond k-anonymity*. ACM Transactions on Knowledge Discovery from Data (TKDD) **1**(1), 3 (2007).

[31] C. Dwork, F. McSherry, K. Nissim, and A. Smith. *Calibrating noise to sensitivity in private data analysis*. In *Theory of Cryptography*, pp. 265–284 (Springer Berlin Heidelberg, 2006).

[32] J. Reed and B. Pierce. *Distance makes the types grow stronger a calculus for differential privacy*. vol. 45, pp. 157–168 (2010).

[33] A. Narayanan and V. Shmatikov. *De-anonymizing social networks*. Proceedings - IEEE Symposium on Security and Privacy (2009).

[34] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. *Privacy: Theory meets practice on the map*. pp. 277–286 (2008).

[35] C. Dwork. *Differential privacy: A survey of results*. In *International conference on theory and applications of models of computation*, pp. 1–19 (Springer, 2008).

[36] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. *What can we learn privately?* In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 531–540 (IEEE, 2008).

[37] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. *Local privacy and statistical minimax rates*. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438 (IEEE, 2013).

[38] U. Erlingsson, V. Pihur, and A. Korolova. *RAPPOR: Randomized aggregatable privacy-preserving ordinal response*. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067 (ACM, New York, NY, USA, 2014).

[39] B. Ding, J. Kulkarni, and S. Yekhanin. *Collecting telemetry data privately*. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3574–3583 (Curran Associates Inc., Red Hook, NY, USA, 2017).

[40] A. Differential Privacy Team. *Learning with privacy at scale*. In *Apple Machine Learning Journal*, vol. 1 (Apple, 2017).

[41] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. *Broadening the scope of differential privacy using metrics*. In *Privacy Enhancing Technologies: 13th International Symposium, PETS 2013, Bloomington, IN, USA, July 10-12, 2013. Proceedings 13*, pp. 82–102 (Springer, 2013).

[42] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. *Geo-indistinguishability: Differential privacy for location-based systems*. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13* (ACM Press, 2013).

[43] A. Falcon. *Aristotle on causality* (2006).

[44] W. C. Salmon. *Causality and explanation* (Oxford University Press, 1998).

[45] M. Bunge. *Causality and modern science* (Routledge, 2017).

[46] J. Pearl. *Causality* (Cambridge university press, 2009).

[47] R. A. Fisher *et al. Statistical methods for research workers.* Statistical methods for research workers. (6th Ed) (1936).

[48] H. S. Bloom. *The core analytics of randomized experiments for social research*. The SAGE handbook of social research methods pp. 115–133 (2008).

[49] D. B. Rubin. *Causal inference using potential outcomes: Design, modeling, decisions*. Journal of the American Statistical Association **100**(469), 322 (2005).

[50] S. L. Morgan and C. Winship. *Counterfactuals and causal inference* (Cambridge University Press, 2015).

[51] B. Schölkopf. *Causality for machine learning*. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804 (2022).

[52] A. P. Dawid. *Statistical causality from a decision-theoretic perspective*. Annual Review of Statistics and Its Application **2**, 273 (2015).

[53] C. Berzuini, P. Dawid, and L. Bernardinell. *Causality: Statistical perspectives and applications* (John Wiley & Sons, 2012).

[54] A. Sjölander. *The language of potential outcomes* (Wiley Online Library, 2012).

[55] I. Shpitser. *Structural equations, graphs and interventions* (Wiley Online Library, 2012).

[56] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference* (Morgan kaufmann, 1988).

[57] A. P. Dawid. *Seeing and doing: The pearlian synthesis*. Heuristics, probability and causality: A tribute to Judea Pearl **309** (2010).

[58] T. J. VanderWeele. *The sufficient cause framework in statistics, philosophy and the biomedical and social sciences* (Wiley Online Library, 2012).

[59] J. L. Mackie. *Causes and conditions*. American philosophical quarterly **2**(4), 245 (1965).

[60] S. Wright. *Correlation and causation* (1921).

[61] T. D. Le, T. Hoang, J. Li, L. Liu, H. Liu, and S. Hu. *A fast pc algorithm for high dimensional causal discovery with multi-core pcs*. IEEE/ACM transactions on computational biology and bioinformatics **16**(5), 1483 (2016).

[62] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search* (MIT press, 2000).

[63] A. Hauser and P. Bühlmann. *Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs*. Journal of Machine Learning Research **13**, 2409 (2012). URL https://jmlr.org/papers/v13/hauser12a.html.

[64] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan. *A linear non-gaussian acyclic model for causal discovery.* Journal of Machine Learning Research **7**(10) (2006).

[65] K. Zhang and A. Hyvarinen. *On the identifiability of the post-nonlinear causal model*. arXiv preprint arXiv:1205.2599 (2012).

[66] D. Janzing and B. Schölkopf. *Causal inference using the algorithmic markov condition*. IEEE Transactions on Information Theory **56**(10), 5168 (2010).

[67] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. *Distinguishing cause from effect using observational data: methods and benchmarks*. The Journal of Machine Learning Research **17**(1), 1103 (2016).

[68] N. Kreif and K. DiazOrdaz. *Machine learning in policy evaluation: new tools for causal inference*. arXiv preprint arXiv:1903.00402 (2019).

[69] R. Aoki and M. Ester. *Causal inference from small high-dimensional datasets*. arXiv preprint arXiv:2205.09281 (2022).

[70] B. K. Lee, J. Lessler, and E. A. Stuart. *Improving propensity score weighting using machine learning*. Statistics in medicine **29**(3), 337 (2010).

[71] C. Tu. *Comparison of various machine learning algorithms for estimating generalized propensity score*. Journal of Statistical Computation and Simulation **89**(4), 708 (2019).

[72] J. Berrevoets, K. Kacprzyk, Z. Qian, and M. van der Schaar. *Causal deep learning*. arXiv preprint arXiv:2303.02186 (2023).

[73] J. Berrevoets, K. Kacprzyk, Z. Qian, and M. van der Schaar. *Navigating causal deep learning*. arXiv preprint arXiv:2212.00911 (2022).

[74] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. *Toward causal representation learning*. Proceedings of the IEEE **109**(5), 612 (2021).

[75] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell. *Representation learning via invariant causal mechanisms*. arXiv preprint arXiv:2010.07922 (2020).

[76] W. Wang, X. Lin, F. Feng, X. He, M. Lin, and T.-S. Chua. *Causal representation learning for out-of-distribution recommendation*. In *Proceedings of the ACM Web Conference 2022*, pp. 3562–3571 (2022).

[77] S. Tople, A. Sharma, and A. Nori. *Alleviating privacy attacks via causal learning*. In *International Conference on Machine Learning*, pp. 9537–9547 (PMLR, 2020).

[78] C. Pinzón, C. Palamidessi, P. Piantanida, and F. Valencia. *On the impossibility of non-trivial accuracy in presence of fairness constraints*. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 7993–8000 (2022).

[79] A. F. Cooper, E. Abrams, and N. Na. *Emergent unfairness in algorithmic fairness-accuracy trade-off research*. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 46–54 (2021).

[80] I. Zliobaite. *On the relation between accuracy and fairness in binary classification*. arXiv preprint arXiv:1505.05723 (2015).

[81] H. Zhao and G. J. Gordon. *Inherent tradeoffs in learning fair representations*. The Journal of Machine Learning Research **23**(1), 2527 (2022).

[82] A. J. London. *Artificial intelligence and black-box medical decisions: accuracy versus explainability*. Hastings Center Report **49**(1), 15 (2019).

[83] S. N. van der Veer, L. Riste, S. Cheraghi-Sohi, D. L. Phipps, M. P. Tully, K. Bozentko, S. Atwood, A. Hubbard, C. Wiper, M. Oswald, *et al. Trading off accuracy and explainability in ai decision-making: findings from 2 citizens' juries*. Journal of the American Medical Informatics Association **28**(10), 2128 (2021).

[84] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua. *Survey on fairness notions and related tensions*. EURO Journal on Decision Processes p. 100033 (2023).

[85] J. S. Kim, J. Chen, and A. Talwalkar. *Fact: A diagnostic for group fairness trade-offs*. In *International Conference on Machine Learning*, pp. 5264–5274 (PMLR, 2020).

[86] L. Xu, C. Jiang, Y. Qian, J. Li, Y. Zhao, and Y. Ren. *Privacy-accuracy trade-off in differentially-private distributed classification: A game theoretical approach*. IEEE Transactions on Big Data **7**(4), 770 (2017).

[87] T. Carvalho, N. Moniz, P. Faria, and L. Antunes. *Towards a data privacy-predictive performance trade-off*. Expert Systems with Applications p. 119785 (2023).

[88] C. Dwork and J. Lei. *Differential privacy and robust statistics*. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380 (2009).

[89] S. B. Hopkins, G. Kamath, M. Majid, and S. Narayanan. *Robustness implies privacy in statistical estimation*. arXiv preprint arXiv:2212.05015 (2022).

[90] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *Certifying and removing disparate impact*. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268 (2015).

[91] T. Calders and S. Verwer. *Three naive bayes approaches for discrimination-free classification*. Data mining and knowledge discovery **21**, 277 (2010).

[92] S. Wei and M. Niethammer. *The fairness-accuracy pareto front*. Statistical Analysis and Data Mining: The ASA Data Science Journal **15**(3), 287 (2022).

[93] Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi. *Understanding and improving fairness-accuracy trade-offs in multi-task learning*. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757 (2021).

[94] S. Beery, G. Van Horn, and P. Perona. *Recognition in terra incognita*. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473 (2018).

[95] A. Rahmattalabi and A. Xiang. *Promises and challenges of causality for ethical machine learning*. arXiv preprint arXiv:2201.10683 (2022).

[96] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. *Fairness through awareness*. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226 (2012).

[97] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning* (fairmlbook.org, 2019). http://www.fairmlbook.org.

[98] F. Kamiran, I. Žliobaitė, and T. Calders. *Quantifying explainable discrimination and removing illegal discrimination in automated decision making*. Knowledge and Information Systems **35**(3), 613 (2013). URL http://link.springer.com/10.1007/s10115-012-0584-8.

[99] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. *Smote: synthetic minority over-sampling technique*. Journal of artificial intelligence research **16**, 321 (2002).

[100] B. Zadrozny and C. Elkan. *Learning and making decisions when costs and probabilities are both unknown*. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 204–213 (2001).

[101] F. Kamiran, T. Calders, and M. Pechenizkiy. *Discrimination aware decision tree learning*. In *2010 IEEE international conference on data mining*, pp. 869–874 (IEEE, 2010).

[102] F. Kamiran and T. Calders. *Classifying without discriminating*. In *2009 2nd international conference on computer, control and communication*, pp. 1–6 (IEEE, 2009).

[103] B. H. Zhang, B. Lemoine, and M. Mitchell. *Mitigating unwanted biases with adversarial learning*. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340 (2018).

[104] L. E. Celis, D. Straszak, and N. K. Vishnoi. *Ranking with fairness constraints*. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)* (Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018).

[105] J. Kleinberg, S. Mullainathan, and M. Raghavan. *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807 (2016).

[106] C. Dwork, A. Roth, *et al. The algorithmic foundations of differential privacy*. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211 (2014).

[107] J. Pearl. *Direct and indirect effects*. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 411–420 (2001).

[108] S. Chiappa. *Path-specific counterfactual fairness*. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 7801–7808 (2019).

[109] Y. Wu, L. Zhang, X. Wu, and H. Tong. *Pc-fairness: A unified framework for measuring causality-based fairness*. In *Advances in Neural Information Processing Systems*, pp. 3404–3414 (2019).

[110] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. *Avoiding discrimination through causal reasoning*. In *Advances in Neural Information Processing Systems*, pp. 656–666 (2017).

[111] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. *Counterfactual fairness*. Advances in neural information processing systems **30** (2017).

[112] D. B. Rubin. *Statistics and causal inference: Comment: Which ifs have causal answers*. Journal of the American Statistical Association **81**(396), 961 (1986).

[113] C. Glymour, K. Zhang, and P. Spirtes. *Review of causal discovery methods based on graphical models*. Frontiers in genetics **10**, 524 (2019).

[114] P. Spirtes and C. Glymour. *An algorithm for fast recovery of sparse causal graphs*. Social science computer review **9**(1), 62 (1991).

[115] D. Entner and P. O. Hoyer. *On causal discovery from time series data using fci*. Probabilistic graphical models pp. 121–128 (2010).

[116] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour. *A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images*. International journal of data science and analytics **3**, 121 (2017).

[117] C. Meek. *Graphical Models: Selecting causal and statistical models*. Ph.D. thesis, Carnegie Mellon University (1997).

[118] D. M. Chickering. *Optimal structure identification with greedy search*. Journal of machine learning research **3**(Nov), 507 (2002).

[119] J. Kuipers, P. Suter, and G. Moffa. *Efficient sampling and structure learning of bayesian networks*. Journal of Computational and Graphical Statistics **31**(3), 639 (2022).

[120] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. *The max-min hill-climbing bayesian network structure learning algorithm*. Machine learning **65**(1), 31 (2006).

[121] P. Blöbaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf. *Cause-effect inference by comparing regression errors*. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909 (PMLR, 2018).

[122] P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. *Inferring deterministic causal relations*. arXiv preprint arXiv:1203.3475 (2012).

[123] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. *Nonlinear causal discovery with additive noise models*. Advances in neural information processing systems **21** (2008).

[124] A. P. Dempster, N. M. Laird, and D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Proceedings of the Royal Statistical Society **B-39**, 1 (1977).

[125] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions* (John Wiley & Sons, 2007).

[126] C. F. J. Wu. *On the convergence properties of the EM algorithm*. The Annals of Statistics **11**(1), 95 (1983).

[127] J. Goodman, O. Gurantz, and J. Smith. *Take two! sat retaking and college enrollment gaps*. American Economic Journal: Economic Policy **12**(2), 115 (2020). URL https://www.aeaweb.org/articles?id=10.1257/pol.20170503.

[128] B. Hannon. *Test anxiety and performance-avoidance goals explain gender differences in sat-v, sat-m, and overall sat scores*. Personality and individual differences **53**, 816 (2012).

[129] M. E. Levine and E. M. Crimmins. *Evidence of accelerated aging among african americans and its implications for mortality*. Social Science & Medicine **118**, 27 (2014).

[130] M. P. Farina, J. K. Kim, and E. M. Crimmins. *Racial/ethnic differences in biological aging and their life course socioeconomic determinants: The 2016 health and retirement study*. Journal of aging and health **35**(3-4), 209 (2023).

[131] G. H. Graf, C. L. Crowe, M. Kothari, D. Kwon, J. J. Manly, I. C. Turney, L. Valeri, and D. W. Belsky. *Testing black-white disparities in biological aging among older adults in the united states: analysis of dna-methylation and blood-chemistry methods*. American journal of epidemiology **191**(4), 613 (2022).

[132] Y. Choi, M. Dang, and G. V. den Broeck. *Group fairness by probabilistic modeling with latent fair decisions*. CoRR **abs/2009.09031** (2020).

[133] R. Islam, S. Pan, and J. R. Foulds. *Fair inference for discrete latent variable models*. arXiv preprint arXiv:2209.07044 (2022).

[134] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. S. Zemel. *The variational fair autoencoder*. In Y. Bengio and Y. LeCun, eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (2016). URL http://arxiv.org/abs/1511.00830.

[135] D. Madras, E. Creager, T. Pitassi, and R. Zemel. *Fairness through causal awareness: Learning causal latent-variable models for biased data*. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 349–358 (2019).

[136] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. *The (im)possibility of fairness: different value systems require different mechanisms for fair decision making*. Commun. ACM **64**(4), 136 (2021).

[137] E. Bareinboim and J. Pearl. *Causal transportability with limited experiments*. In *AAAI* (2013).

[138] E. Bareinboim and J. Pearl. *A general algorithm for deciding transportability of experimental results*. Journal of Causal Inference **1**(1), 107 (2013). URL https://doi.org/10.1515%2Fjci-2012-0004.

[139] E. Bareinboim and J. Pearl. *Transportability from multiple environments with limited experiments: Completeness results*. Advances in neural information processing systems **27** (2014).

[140] J. Pearl and E. Bareinboim. *External validity: From do-calculus to transportability across populations*. Statistical Science **29**(4) (2014). URL https://doi.org/10.1214%2F14-sts486.

[141] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. *Towards causal representation learning*. CoRR **abs/2102.11107** (2021).

[142] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. *Algorithmic decision making and the cost of fairness*. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806 (2017).

[143] M. Glymour, J. Pearl, and N. P. Jewell. *Causal inference in statistics: A primer* (John Wiley & Sons, 2016).

[144] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. *Causal effect inference with deep latent-variable models*. Advances in neural information processing systems **30** (2017).

[145] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *Certifying and removing disparate impact*. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, p. 259–268 (Association for Computing Machinery, New York, NY, USA, 2015). URL https://doi.org/10.1145/2783258.2783311.

[146] D. Agrawal and C. C. Aggarwal. *On the design and quantification of privacy preserving data mining algorithms*. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '01, p. 247–255 (Association for Computing Machinery, New York, NY, USA, 2001). URL https://doi.org/10.1145/375551.375602.

[147] E. ElSalamouny and C. Palamidessi. *Generalized iterative bayesian update and applications to mechanisms for privacy protection*. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 490–507 (IEEE, 2020).

[148] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, *et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*. IBM Journal of Research and Development **63**(4/5), 4 (2019).

[149] D. Kwon and D. W. Belsky. *A toolkit for quantification of biological age from blood chemistry and organ function test data: Bioage*. GeroScience **43**, 2795 (2021).

[150] Y. Xu, X. Wang, D. W. Belsky, W. V. McCall, Y. Liu, and S. Su. *Blunted rest–activity circadian rhythm is associated with increased rate of biological aging: an analysis of nhanes 2011–2014*. The Journals of Gerontology: Series A **78**(3), 407 (2023).

[151] W. Liu, J. Wang, M. Wang, H. Hou, X. Ding, L. Ma, and M. Liu. *Oxidative stress factors mediate the association between life's essential 8 and accelerated phenotypic aging: Nhanes 2005-2018*. The Journals of Gerontology: Series A p. glad240 (2023).

[152] L. Nguyen, J. Chon, E. Kim, J. Cheng, and J. Ebersole. *Biological aging and periodontal disease: analysis of nhanes (2001–2002)*. JDR Clinical & Translational Research **7**(2), 145 (2022).

[153] I. Chen, F. D. Johansson, and D. Sontag. *Why is my classifier discriminatory?* Advances in neural information processing systems **31** (2018).

[154] P. Domingos. *A unified bias-variance decomposition*. In *Proceedings of 17th international conference on machine learning*, pp. 231–238 (Morgan Kaufmann Stanford, 2000).

[155] K. Karkkainen and J. Joo. *Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation*. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1548–1558 (2021).

[156] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky. *Fair attribute classification through latent space de-biasing*. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9301–9310 (2021).

[157] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, and D. Tzovaras. *Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems*. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2302–2314 (2022).

[158] S. Yucer, S. Akçay, N. Al-Moubayed, and T. P. Breckon. *Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 18–19 (2020).

[159] Y. Zhang and J. Sang. *Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing*. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 4346–4354 (2020).

[160] T. Xu, J. White, S. Kalkan, and H. Gunes. *Investigating bias and fairness in facial expression recognition*. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 506–523 (Springer, 2020).

[161] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. *Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations*. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 5309–5318 (2018).

[162] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. *Towards fairness in visual recognition: Effective strategies for bias mitigation*. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8919–8928 (2020).

[163] M. Qraitem, K. Saenko, and B. A. Plummer. *Bias mimicking: A simple sampling approach for bias mitigation*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20320 (2023).

[164] H. He and E. A. Garcia. *Learning from imbalanced data*. IEEE Transactions on knowledge and data engineering **21**(9), 1263 (2009).

[165] R. Kohavi, D. H. Wolpert, *et al. Bias plus variance decomposition for zero-one loss functions*. In *ICML*, vol. 96, pp. 275–83 (Citeseer, 1996).

[166] R. Kohavi *et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.* In *Kdd*, vol. 96, pp. 202–207 (1996).

[167] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. *Machine bias. propublica*. See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).

[168] E. S. Nordholt, M. Hartgers, and R. Gircour. *The dutch virtual census of 2001*. Analysis and Methodology (2004).

[169] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, *et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*. IBM Journal of Research and Development **63**(4/5), 4 (2019).

[170] T. Farrand, F. Mireshghallah, S. Singh, and A. Trask. *Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy*. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, pp. 15–19 (2020).

[171] S. M. Lundberg and S. Lee. *A unified approach to interpreting model predictions*. In *NIPS*, pp. 4765–4774 (2017).

[172] A. R. Nogueira, J. Gama, and C. A. Ferreira. *Causal discovery in machine learning: Theories and applications*. Journal of Dynamics & Games **8**(3), 203 (2021).

[173] J. Jordon, J. Yoon, and M. Van Der Schaar. *Pate-gan: Generating synthetic data with differential privacy guarantees*. In *International conference on learning representations* (2019).

[174] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. *Privbayes: Private data release via bayesian networks*. ACM Trans. Database Syst. **42**(4) (2017). URL https://doi.org/10.1145/3134428.

[175] P. Kairouz, K. Bonawitz, and D. Ramage. *Discrete distribution estimation under local privacy*. In *Int. Conf. on Machine Learning*, pp. 2436–2444 (PMLR, 2016).

[176] M. J. Kusner, Y. Sun, K. Sridharan, and K. Q. Weinberger. *Private causal inference*. In A. Gretton and C. C. Robert, eds., *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51 of *Proceedings of Machine Learning Research*, pp. 1308–1317 (PMLR, Cadiz, Spain, 2016).

[177] D. Xu, S. Yuan, and X. Wu. *Differential privacy preserving causal graph discovery*. In *2017 IEEE Symposium on Privacy-Aware Computing (PAC)*, pp. 60–71 (2017).

[178] L. Wang, Q. Pang, and D. Song. *Towards practical differentially private causal graph discovery*. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., *Advances in Neural Information Processing Systems*, vol. 33, pp. 5516–5526 (Curran Associates, Inc., 2020). URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3b13b1eb44b05f57735764786fab9c2c-Paper.pdf.

[179] P. Ma, Z. Ji, Q. Pang, and S. Wang. *Noleaks: Differentially private causal discovery under functional causal model*. IEEE Transactions on Information Forensics and Security **17**, 2324 (2022).

[180] T. Wang, J. Blocki, N. Li, and S. Jha. *Locally differentially private protocols for frequency estimation*. In *26th USENIX Security Symposium (USENIX Security 17)*, pp. 729–745 (USENIX Association, Vancouver, BC, 2017).

[181] J. Acharya, Z. Sun, and H. Zhang. *Hadamard response: Estimating distributions privately, efficiently, and with little communication*. In K. Chaudhuri and M. Sugiyama, eds., *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, vol. 89 of *Proceedings of Machine Learning Research*, pp. 1120–1129 (PMLR, 2019).

[182] H. H. Arcolezi, J.-F. Couchot, B. A. Bouna, and X. Xiao. *Improving the utility of locally differentially private protocols for longitudinal and multidimensional frequency estimates*. Digital Communications and Networks (2022).

[183] H. Kikuchi. *Castell: Scalable joint probability estimation of multi-dimensional data randomized with local differential privacy*. arXiv preprint arXiv:2212.01627 (2022).

[184] G. Cormode, S. Jha, T. Kulkarni, N. Li, D. Srivastava, and T. Wang. *Privacy at scale: Local differential privacy in practice*. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 1655–1658 (2018).

[185] A. Agarwal and R. Singh. *Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy*. arXiv preprint arXiv:2107.02780 (2021).

[186] Y. Ohnishi and J. Awan. *Locally private causal inference*. arXiv preprint arXiv:2301.01616 (2023).

[187] S. L. Warner. *Randomized response: A survey technique for eliminating evasive answer bias*. Journal of the American Statistical Association **60**(309), 63 (1965).

[188] J. Domingo-Ferrer and J. Soria-Comas. *Multi-dimensional randomized response*. IEEE Transactions on Knowledge and Data Engineering **34**(10), 4933 (2022).

[189] H. H. Arcolezi, S. Gambs, J.-F. Couchot, and C. Palamidessi. *On the risks of collecting multidimensional data under local differential privacy*. Proc. VLDB Endow. **16**(5), 1126–1139 (2023).

[190] K. Chatzikokolakis, G. Cherubin, C. Palamidessi, and C. Troncoso. *Bayes security: A not so average metric*. In *2023 2023 IEEE 36th Computer Security Foundations Symposium (CSF) (CSF)*, pp. 159–177 (IEEE Computer Society, Los Alamitos, CA, USA, 2023).

[191] F. L. Rios, G. Moffa, and J. Kuipers. *Benchpress: a scalable and versatile workflow for benchmarking structure learning algorithms for graphical models* (2021). 2107.03863.

[192] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. *Causal protein-signaling networks derived from multiparameter single-cell data*. Science **308**(5721), 523 (2005).

[193] H. Han, Y. Ma, and W. Zhu. *Galton's family heights data revisited*. arXiv preprint arXiv:1508.02942 (2015).

[194] R. C. Johnson, G. E. McClearn, S. Yuen, C. T. Nagoshi, F. M. Ahern, and R. E. Cole. *Galton's data a century later.* American Psychologist **40**(8), 875 (1985).

[195] D. Kalainathan, O. Goudet, and R. Dutta. *Causal discovery toolbox: Uncovering causal relationships in python*. The Journal of Machine Learning Research **21**(1), 1406 (2020).

[196] J. A. Fonollosa. *Conditional distribution variability measures for causality detection*. Cause Effect Pairs in Machine Learning pp. 339–347 (2019).

[197] C. Pinzón, C. Rocha, and J. Finke. *An approach to optimal discretization of continuous real random variables with application to machine learning* (2020).

[198] C. J. Clopper and E. S. Pearson. *The use of confidence or fiducial limits illustrated in the case of the binomial*. Biometrika **26**(4), 404 (1934).

[199] J. Zhang and E. Bareinboim. *Fairness in decision-making—the causal explanation formula*. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018).

[200] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. *Interventional fairness: Causal database repair for algorithmic fairness*. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 793–810 (2019).

[201] A. Khademi, S. Lee, D. Foley, and V. Honavar. *Fairness in algorithmic decision making: An excursion through the lens of causality*. In *The World Wide Web Conference*, pp. 2907–2914 (2019).

[202] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. *Causal reasoning for algorithmic fairness*. arXiv preprint arXiv:1805.05859 (2018).

[203] A. Kasirzadeh and A. Smart. *The use and misuse of counterfactuals in ethical machine learning* (2021). URL https://arxiv.org/abs/2102.05085.

[204] N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, and R. Silva. *The sensitivity of counterfactual fairness to unmeasured confounding*. In *Uncertainty in artificial intelligence*, pp. 616–626 (PMLR, 2020).

[205] J. Fawkes, R. Evans, and D. Sejdinovic. *Selection, ignorability and challenges with causal fairness*. In *Conference on Causal Learning and Reasoning*, pp. 275–289 (PMLR, 2022).

[206] A. Xiang. *Reconciling legal and technical approaches to algorithmic bias*. Tenn. L. Rev. **88**, 649 (2020).

[207] A. Xiang and I. D. Raji. *On the legal compatibility of fairness definitions*. arXiv preprint arXiv:1912.00761 (2019).

[208] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva. *Causal reasoning for algorithmic fairness*. arXiv preprint arXiv:1805.05859 (2018).

[209] S. Haack. *Evidence matters: Science, proof, and truth in the law* (Cambridge University Press, 2014).

[210] L. B. Solum. *Procedural justice*. S. CAl. l. reV. **78**, 181 (2004).

[211] N. Rescher. *Plausible reasoning: An introduction to the theory and practice of plausibilistic inference*. Philosophy and Rhetoric **13**(3) (1980).

[212] P. Spirtes, C. Meek, and T. Richardson. *An algorithm for causal inference in the presence of latent variables and selection bias*. Computation, causation, and discovery **21**, 211 (1999).

[213] I. Shpitser and J. Pearl. *What counterfactuals can be tested*. In *23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007*, pp. 352–359 (2007).

[214] I. Shpitser and J. Pearl. *Complete identification methods for the causal hierarchy*. Journal of Machine Learning Research **9**(Sep), 1941 (2008).

[215] L. Grozdanovski. *In search of effectiveness and fairness in proving algorithmic discrimination in eu law*. Common Market Law Review **58**(1) (2021).

[216] T. J. VanderWeele and M. A. Hernán. *Causal effects and natural laws: towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex*. Causality: statistical perspectives and applications pp. 101–113 (2012).

[217] P. W. Holland. *Statistics and causal inference*. Journal of the American statistical Association **81**(396), 945 (1986).

[218] D. Neumark. *Experimental research on labor market discrimination*. Journal of Economic Literature **56**(3), 799 (2018).

[219] L. Keele. *The statistics of causal inference: A view from political methodology*. Political Analysis **23**(3), 313 (2015).

[220] L. Laffers and G. Mellace. *Identification of the average treatment effect when sutva is violated*. Discussion Papers on Business and Economics, University of Southern Denmark **3** (2020).

[221] D. Westreich and S. R. Cole. *Invited commentary: positivity in practice*. American journal of epidemiology **171**(6), 674 (2010).

[222] T. VanderWeele and S. Vansteelandt. *Mediation analysis with multiple mediators*. Epidemiologic methods **2**(1), 95 (2014).

[223] R. Binkytė-Sadauskienė, K. Makhlouf, C. Pinzón, S. Zhioua, and C. Palamidessi. *Causal discovery for fairness*. arXiv preprint arXiv:2206.06685 (2022).

[224] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J. P. Pellet, P. Spirtes, and A. Statnikov. *Causality workbench*. In *Causality in the sciences* (Oxford University Press, 2011).

[225] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences* (Cambridge University Press, 2015).

[226] H. Cramér. *Mathematical methods of statistics*, vol. 43 (Princeton university press, 1999).

[227] J. Pearl. *Linear models: A useful "microscope" for causal analysis*. Journal of Causal Inference **1**(1), 155 (2013).

[228] J. Rawls. *A theory of justice: Revised edition* (Harvard university press, 2020).

[229] J. Pearl. *On measurement bias in causal inference*. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 425–432 (2010).

[230] C. R. Weinberg. *Can dags clarify effect modification?* Epidemiology (Cambridge, Mass.) **18**(5), 569 (2007).

[231] K. J. Rothman, S. Greenland, T. L. Lash, *et al. Modern epidemiology*, vol. 3 (Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008).

[232] T. J. VanderWeele and J. M. Robins. *The identification of synergism in the sufficient-component-cause framework*. Epidemiology **18**(3), 329 (2007).

[233] L. Keele and R. T. Stevenson. *Causal interaction and effect modification: same model, different concepts*. Political Science Research and Methods **9**(3), 641–649 (2021).

[234] T. J. VanderWeele. *Controlled direct and mediated effects: definition, identification and bounds*. Scandinavian Journal of Statistics **38**(3), 551 (2011).

[235] E. H. Simpson. *The interpretation of interaction in contingency tables*. Journal of the Royal Statistical Society: Series B (Methodological) **13**(2), 238 (1951).

[236] S. Haneuse. *Distinguishing selection bias and confounding bias in comparative effectiveness research*. Medical care **54**(4), e23 (2016).

[237] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. *A survey on datasets for fairness-aware machine learning*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), e1452 (2022).

[238] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. *Machine bias*. In *Ethics of data and analytics*, pp. 254–264 (Auerbach Publications, 2022).

[239] L. Zhang, Y. Wu, and X. Wu. *Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms*. IEEE Transactions on Knowledge and Data Engineering **31**(11), 2035 (2018).

[240] W. Huan, Y. Wu, L. Zhang, and X. Wu. *Fairness through equality of effort*. In *Companion Proceedings of the Web Conference 2020*, pp. 743–751 (2020).

[241] S. B. Ahmed and S. M. Dumanski. *Sex, gender and COVID-19: a call to action*. Canadian Journal of Public Health **111**(6), 980 (2020).

[242] C. Gebhard, V. Regitz-Zagrosek, H. K. Neuhauser, R. Morgan, and S. L. Klein. *Impact of sex and gender on COVID-19 outcomes in Europe*. Biology of Sex Differences **11**(1), 1 (2020).

[243] T. Smith. *A Supercomputer Analyzed COVID-19—and an Interesting New Theory Has Emerged* (2020). Elemental https://elemental.medium.com/a-supercomputer-analyzed-covid-19-and-an-interesting-new-theory-has-emerged-31cb8eba9d63.

[244] S. L. Klein, S. Dhakal, R. L. Ursin, S. Deshpande, K. Sandberg, and F. Mauvais-Jarvis. *Biological sex impacts COVID-19 outcomes*. PLoS pathogens **16**(6), e1008570 (2020).

[245] C. P. Tadiri, T. Gisinger, A. Kautzky-Willer, K. Kublickiene, M. T. Herrero, V. Raparelli, L. Pilote, and C. M. Norris. *The influence of sex and gender domains on covid-19 cases and mortality*. Cmaj **192**(36), E1041 (2020).

[246] J. Kopel, A. Perisetti, A. Roghani, M. Aziz, M. Gajendran, and H. Goyal. *Racial and gender-based differences in COVID-19*. Frontiers in public health **8**, 418 (2020).

[247] M. Besserve, S. Buchholz, and B. Schölkopf. *Assaying Large-scale Testing Models to Interpret COVID-19 Case Numbers* (2021). 2012.01912.

[248] S. L. Klein, A. Jedlicka, and A. Pekosz. *The Xs and Y of immune responses to viral vaccines*. The Lancet infectious diseases **10**(5), 338 (2010).

[249] S. E. Chiarella, C. Pabelick, and Y. Prakash. *Sex differences in the coronavirus disease 2019*. In *Sex-Based Differences in Lung Physiology*, pp. 471–490 (Springer, 2021).

[250] G. Sharma, A. S. Volgman, and E. D. Michos. *Sex differences in mortality from COVID-19 pandemic: are men vulnerable and women protected?* Case Reports **2**(9), 1407 (2020).

[251] H. Peckham, N. M. de Gruijter, C. Raine, A. Radziszewska, C. Ciurtin, L. R. Wedderburn, E. C. Rosser, K. Webb, and C. T. Deakin. *Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission*. Nature communications **11**(1), 1 (2020).

[252] A. M. Traish and A. Morgentaler. *What's testosterone got to do with it? a critical assessment of the contribution of testosterone to gender disparities in covid-19 infections and deaths*. Androgens: Clinical Research and Therapeutics **2**(1), 18 (2021).

[253] L. Grzelak, A. Velay, Y. Madec, F. Gallais, I. Staropoli, C. Schmidt-Mutter, M.-J. Wendling, N. Meyer, C. Planchais, D. Rey, *et al. Sex differences in the decline of neutralizing antibodies to SARS-CoV-2*. medRxiv (2020).

[254] C. Wenham, J. Smith, and R. Morgan. *COVID-19: the gendered impacts of the outbreak*. The Lancet **395**(10227), 846 (2020).

[255] J. von Kügelgen, L. Gresele, and B. Schölkopf. *Simpson's paradox in COVID-19 case fatality rates: a mediation analysis of age-related causal effects*. arXiv preprint arXiv:2005.07180 (2020).

[256] J. R. Head, K. Andrejko, Q. Cheng, P. A. Collender, S. Phillips, A. Boser, A. K. Heaney, C. M. Hoover, S. L. Wu, G. R. Northrup, *et al. The effect of school closures and reopening strategies on COVID-19 infection dynamics in the San Francisco Bay Area: a cross-sectional survey and modeling analysis*. medRxiv (2020).

[257] D. Bertsimas, L. Boussioux, R. C. Wright, A. Delarue, V. Digalakis Jr, A. Jacquillat, D. L. Kitane, G. Lukin, M. L. Li, L. Mingardi, *et al. From predictions to prescriptions: A data-driven response to COVID-19*. arXiv preprint arXiv:2006.16509 (2020).

[258] R. De La Vega, R. Ruíz-Barquín, S. Boros, and A. Szabo. *Could attitudes toward covid-19 in spain render men more vulnerable than women?* Global public health **15**(9), 1278 (2020).

[259] G. M. Bwire. *Coronavirus: Why Men are More Vulnerable to COVID-19 Than Women?* Sn Comprehensive Clinical Medicine p. 1 (2020).

[260] H. Kabir, M. Maple, K. Usher, and M. S. Islam. *Health vulnerabilities of readymade garment (RMG) workers: a systematic review*. BMC Public Health **19**(1), 1 (2019).

[261] P. Silpasuwan, S. Prayomyong, D. Sujitrat, and P. Suwan-Ampai. *Cotton dust exposure and resulting respiratory disorders among home-based garment workers*. Workplace Health & Safety **64**(3), 95 (2016).

[262] P. M. Dana, F. Sadoughi, J. Hallajzadeh, Z. Asemi, M. A. Mansournia, B. Yousefi, and M. Momen-Heravi. *An insight into the sex differences in covid-19 patients: what are the possible causes?* Prehospital and disaster medicine **35**(4), 438 (2020).

[263] M. Zietz, J. Zucker, and N. P. Tatonetti. *Associations between blood type and COVID-19 infection, intubation, and death*. Nature communications **11**(1), 1 (2020).

[264] M. Zietz and N. P. Tatonetti. *Testing the association between blood type and COVID-19 infection, intubation, and death*. MedRxiv (2020).

[265] F. Pourali, M. Afshari, R. Alizadeh-Navaei, J. Javidnia, M. Moosazadeh, and A. Hessami. *Relationship between blood group and risk of infection and death in COVID-19: a live meta-analysis*. New Microbes and New Infections **37**, 100743 (2020). URL https://www.sciencedirect.com/science/article/pii/S2052297520300950.

[266] M. Ebadi and A. J. Montano-Loza. *Perspective: improving vitamin D status in the management of COVID-19*. European journal of clinical nutrition **74**(6), 856 (2020).

[267] A. Jain, R. Chaurasia, N. S. Sengar, M. Singh, S. Mahor, and S. Narain. *Analysis of vitamin D level among asymptomatic and critically ill COVID-19 patients and its correlation with inflammatory markers*. Scientific reports **10**(1), 1 (2020).

[268] H. Zeberg and S. Pääbo. *The major genetic risk factor for severe COVID-19 is inherited from Neanderthals*. Nature **587**(7835), 610 (2020).

[269] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. *Learning functional causal models with generative neural networks*. In *Explainable and interpretable models in computer vision and machine learning*, pp. 39–80 (Springer, 2018).

[270] J. Pearl and R. Dechter. *Identifying independencies in causal graphs with feedback*. arXiv preprint arXiv:1302.3595 (2013).

[271] D. Geiger and J. Pearl. *On the logic of causal models*. In *Machine Intelligence and Pattern Recognition*, vol. 9, pp. 3–14 (Elsevier, 1990).

[272] L. Starkopf, M. Andersen, T. Gerds, C. Torp-Pedersen, and T. Lange. *Comparison of five software solutions to mediation analysis*. Copenhagen, Denmark: University of Copenhagen (2017).

[273] X. Zhou and T. Yamamoto. *Tracing causal paths from experimental and observational data* (2020). URL osf.io/preprints/socarxiv/2rx6p.

[274] D. Tingley, T. Yamamoto, K. Hirose, L. Keele, and K. Imai. *Mediation: R package for causal mediation analysis*. Journal of Statistical Software (2014).

[275] J. Pearl. *Interpretation and identification of causal mediation.* Psychological methods **19**(4), 459 (2014).

[276] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. *Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment*. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180 (2017).

[277] K. Power. *The COVID-19 pandemic has increased the care burden of women and families*. Sustainability: Science, Practice and Policy **16**(1), 67 (2020).

A

Appendix to Chapter 7

A.0.1 Additional plots for the experiments on synthetic data described in the body of the paper

We recall that the sequence of values s_i for the binary sensitive variable (group) S are generated by sampling from the Bernoulli distribution $\mathcal{B}(0.5)$. The domain of E is set to be equal to $[0, 99]$, and to each of the elements s_i in the data set we associate a value e_i for the variable E, sampled from the normal distribution $\mathcal{N}(mean1, sd)$ if $s_i = 1$ and from $\mathcal{N}(mean0, sd)$ if $s_i = 0$. The mean *mean1* is set to be 60, while the value of *mean0* varies through the experiments from 40 to 80. The standard deviation sd is set to be 30. We keep the samples in the range of E by re-sampling the values that are lower than 0 or higher than 99. We also discretize them by rounding to the nearest integer. Finally, to each pair (s_i, e_i) we associate a value z_i for Z by applying a bias to e_i with a certain probability. More precisely, $z_i = e_i + (bias \times e_i)$, where *bias* is sampled from $\mathcal{N}(-0.2, 0.05)$ (negative bias) if $s_i = 0$ and from $\mathcal{N}(0.2, 0.05)$ (positive bias) if $s_i = 1$.

The boxplots for the various metrics are obtained by repeating the experiments 10 times, with different sampling from the same original distributions.

The difference between the original and the estimated distributions is measured using the Wasserstein distance. The results, for each group separately, and the two groups combined, are shown in Figures A.1, A.2, and A.3 respectively.

We now compute by BaBE the empirical distributions $\hat{P}[E|Z, S]$. We verify that these satisfy the conditions for Method 1, and we apply this method to set the values of $\hat{E}$ and $\hat{Y}_{\hat{E}}$ for each sample. Based on this, we compute various metrics for precision and fairness, and compare them with the results obtained with the methods DI and NB. We also compare them with the prediction based on Z, namely $\hat{Y}_{\hat{Z}}$. We recall that the threshold for the decision is $E = 60$. Namely, $Y_E = 1$ if $E > 60$ and $Y_E = 0$ otherwise. The threshold is the same for Z, i.e., $Y_Z = 1$ if $Z > 60$ and $Y_Z = 0$

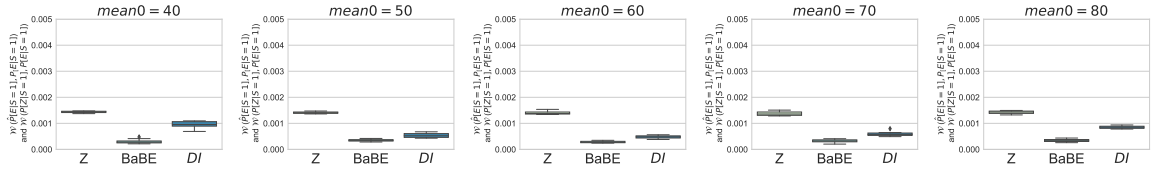


Figure A.1: The Wasserstein distance between $\hat{P}[Z|S=1]$ and $P[E|S=1]$ and between $\hat{P}[E|S=1]$ and $P[E|S=1]$

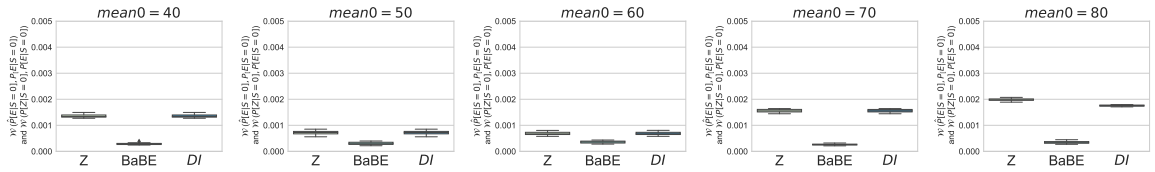


Figure A.2: The Wasserstein distance between $\hat{P}[Z|S=0]$ and $P[E|S=0]$ and between $\hat{P}[E|S=0]$ and $P[E|S=0]$

otherwise.

Figures A.4, A.4 and A.6 show the accuracy of a prediction based on Z (with respect to the true decision based on E), and the accuracy of the predictions obtained with the BaBE, DI, and NB methods. We show the accuracy obtained separately for each group, and then for the two groups combined.

Next, we show in Figures A.7 and A.8 the probabilities of positive prediction on admission for each group, conditioned on $E = 55$. Note that, because of the positive bias that we have for group 1 in the data, the prediction based on Z is positive with a high probability for group 0, whereas the true decision (one based on E) should be negative because the threshold is $E = 60$. The prediction based on the E estimated by BaBE, on the other hand, is correct, in the sense that the probability of a positive prediction is very small. For group 0 the bias is negative, hence the prediction is negative also when is based on Z. Figure A.9 shows the Conditional Statistical Parity Difference based obtained from these probabilities.

Finally, Figures A.10, A.11, and A.12 show the probabilities of positive prediction when the true decision is positive, and the corresponding Equal Opportunity Difference. Note that the prediction based on Z has high probability to be positive for group 1, but not for group 0. This is due to the fact that the Z for group 1 has a positive bias w.r.t. E, while for group 0 the bias is negative. Hence, for group 1, whenever E is greater than the threshold ($E \geq 60$), also Z is very likely to be greater, while this is not the case for group 0. On the other hand, BaBE's prediction is based on the estimation of E, and hence tends to be equal to the true decision.

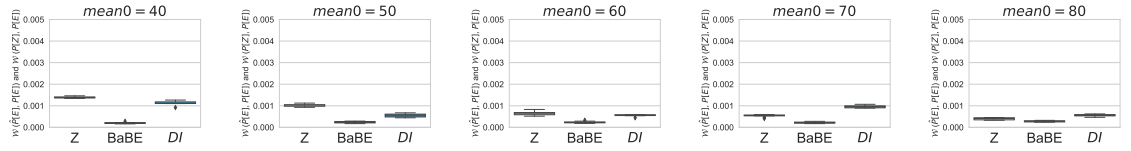


Figure A.3: The Wasserstein distance between $\hat{P}[Z]$ and $P[E]$ and between $\hat{P}[E]$ and $P[E]$

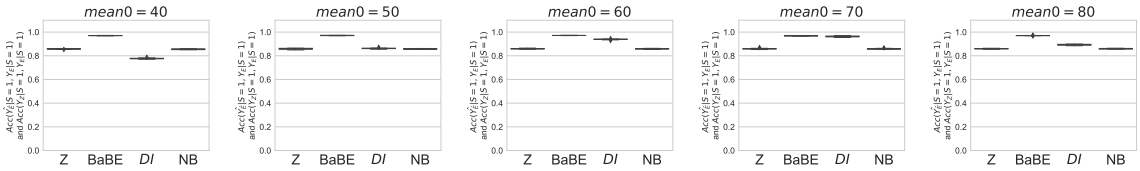


Figure A.4: The accuracy of $\hat{Y}_Z|S=1$ and $\hat{Y}_{\hat{E}}|S=1$ w.r.t. $Y_E|S=1$ (for Z).

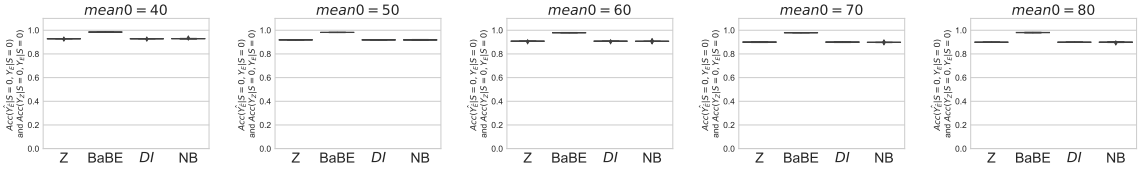


Figure A.5: The accuracy between of $\hat{Y}_Z|S=0$ and $\hat{Y}_{\hat{E}}|S=0$ w.r.t. $Y_E|S=0$.

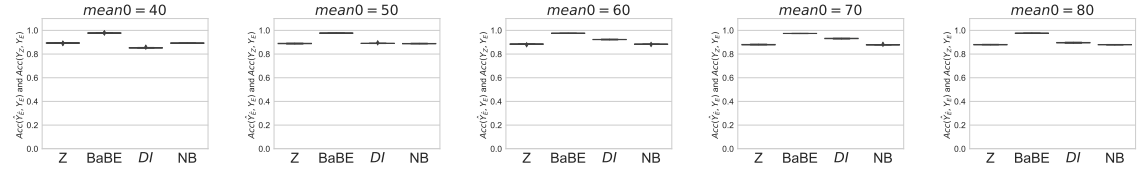


Figure A.6: The accuracy of $\hat{Y}_Z$ and $\hat{Y}_{\hat{E}}$ and Y_E w.r.t. Y_E (for Z).

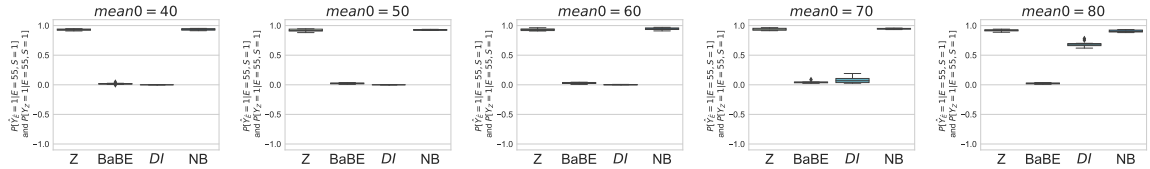


Figure A.7: $P[Y_Z=1|E=55,S=1]$ and $P[\hat{Y}_{\hat{E}}=1|E=55,S=1]$.

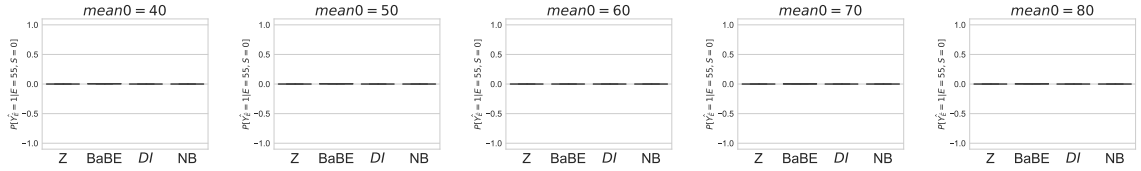


Figure A.8: $P[Y_Z = 1|E = 55, S = 0]$ and $P[\hat{Y}_{\hat{E}} = 1|E = 55, S = 0]$.

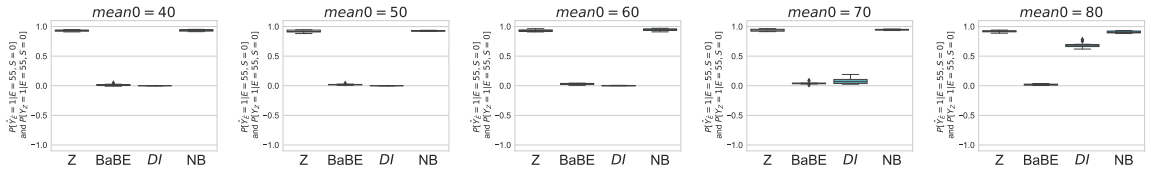


Figure A.9: Conditional Statistical Parity Difference (CSPD_{55}). We recall that for BaBE, DI and NB, CSPD_{55} is defined as $P[\hat{Y}_{\hat{E}} = 1|E = 55, S = 1] - P[\hat{Y}_{\hat{E}} = 1|E = 55, S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z.

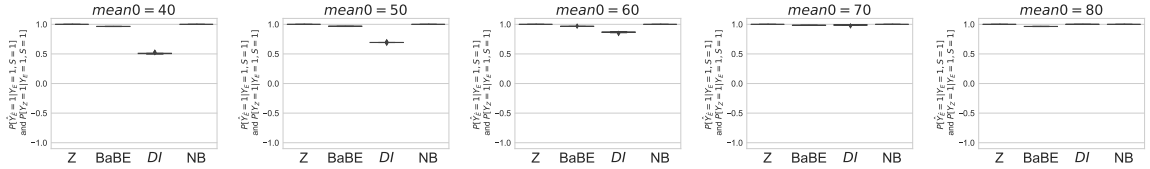


Figure A.10: $P[Y_Z = 1|Y_E = 1, S = 1]$ and $P[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1]$.

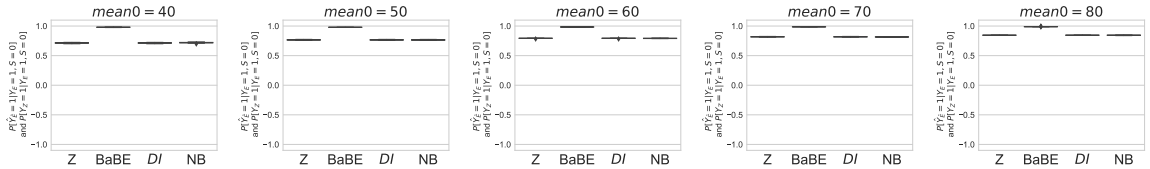


Figure A.11: $P[Y_Z = 1|Y_E = 1, S = 0]$ and $P[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 0]$.

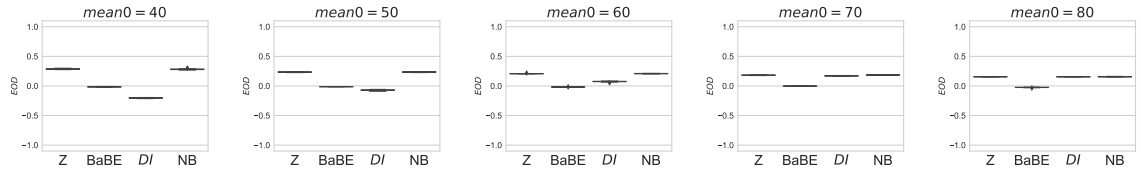


Figure A.12: Equal Opportunity Difference (EOD). We recall that for BaBE, DI and NB, EOD is defined as $P[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 1] - P[\hat{Y}_{\hat{E}} = 1|Y_E = 1, S = 0]$. For Z, the definition is similar, with $\hat{Y}_{\hat{E}}$ replaced by Y_Z.

B

Appendix to Chapter 8

B.0.1 Additional plots for the magnitude of SSB and URB (Sections 8.5.1 and 8.5.2)

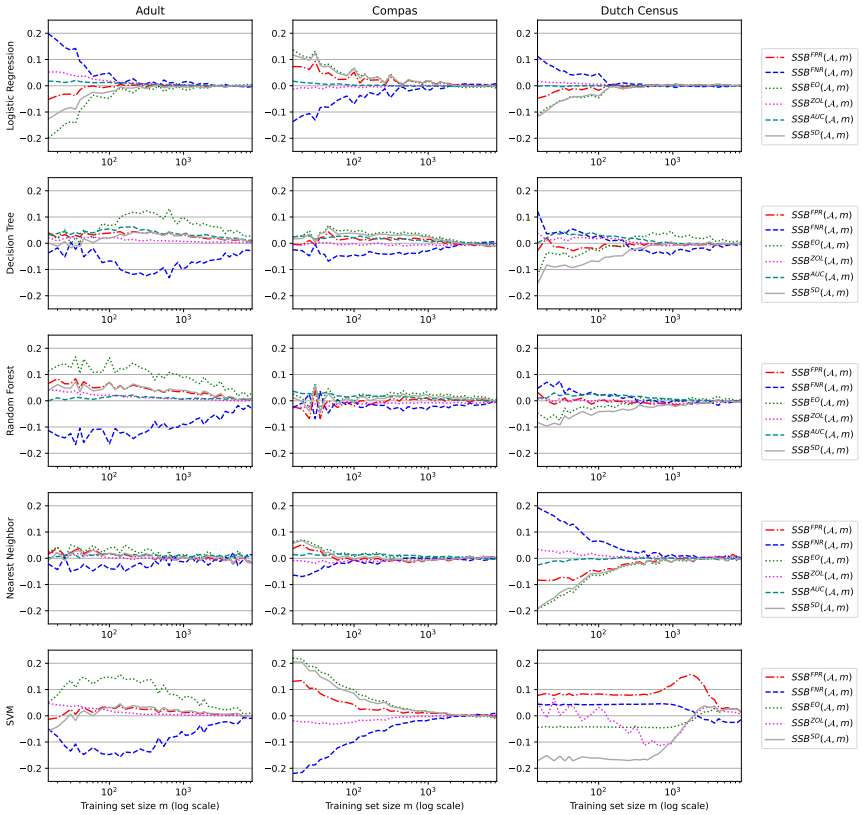


Figure B.1: Additional plots for the magnitude of SSB and URB . Magnitude of sample size bias (SSB) for increasing size of the training data.

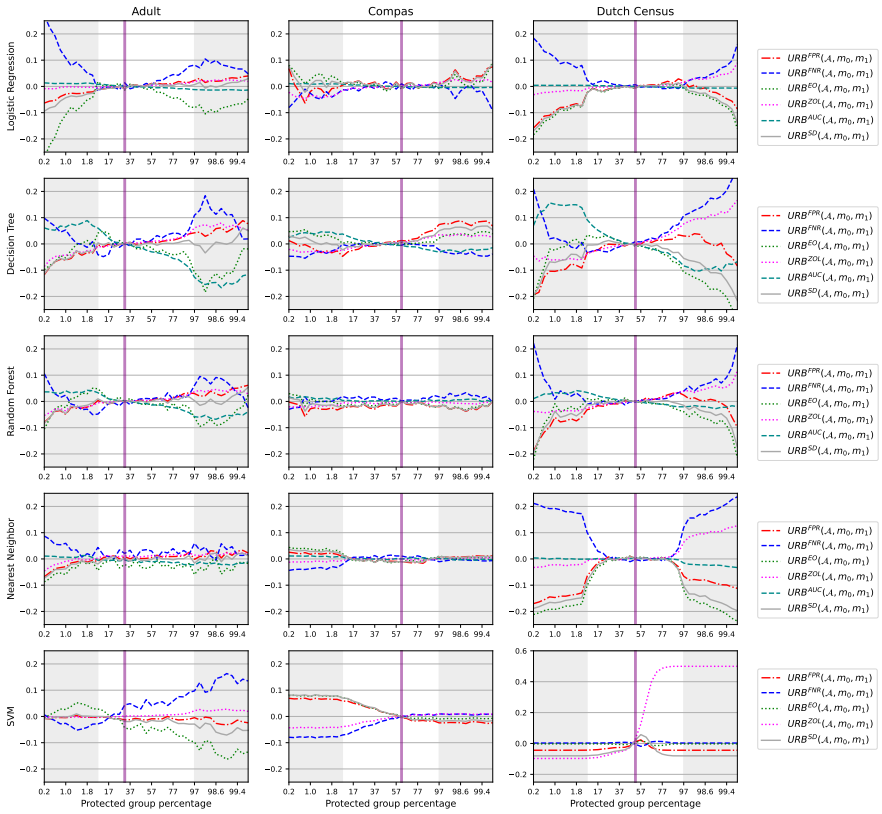


Figure B.2: Additional plots for the magnitude of SSB and URB. Underrepresentation Bias (URB) for different ratios of sensitive groups. The training set size is fixed (1000). The horizontal bar represents the same ratio as the population. The shaded sections indicate a focus on the extreme proportions (less than 2% and more than 98%).

B.0.2 Additional plots for the effect of collecting more samples on discrimination (Section 8.5.4)

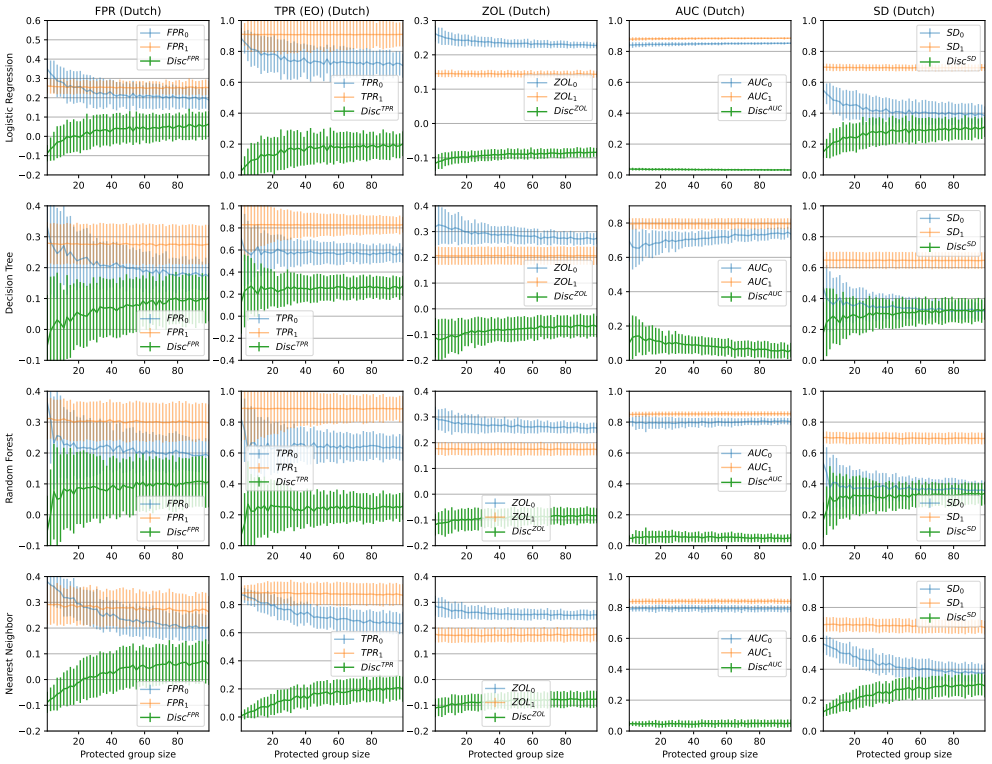


Figure B.3: Additional plots for the effect of collecting more samples on discrimination. Discrimination values for the Dutch Census dataset while increasing the size of the protected group.

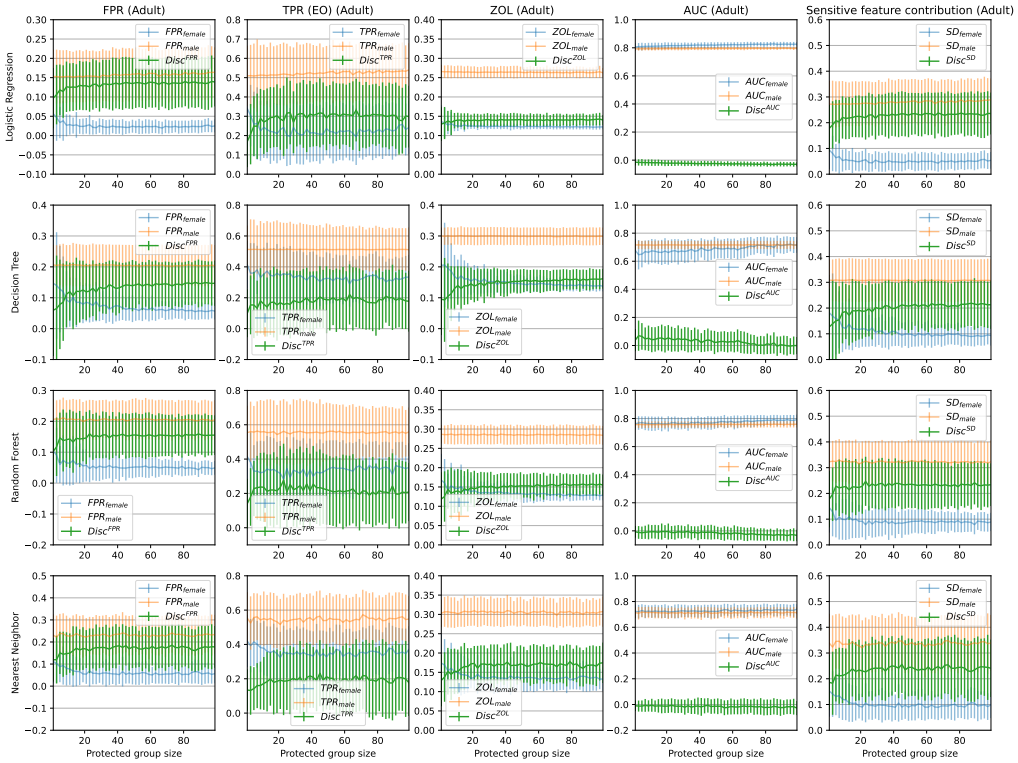


Figure B.4: Additional plots for the effect of collecting more samples on discrimination. Discrimination value for the Adult dataset while increasing the size of the protected group.

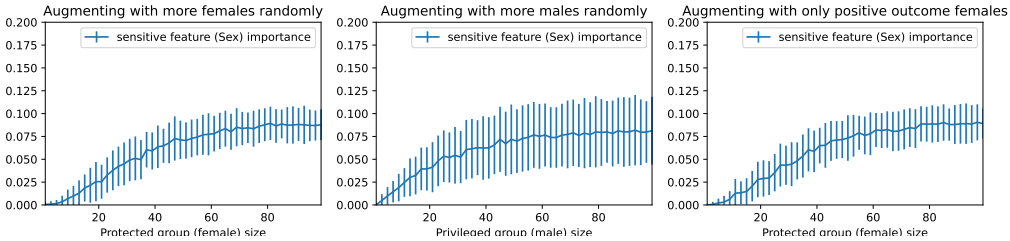


Figure B.5: Additional plots for the effect of collecting more samples on discrimination. Sensitive feature (Sex) importance observed in the experiments of Section 8.5.4.

C

Appendix to Chapter 9

C.1 Privacy Mechanisms

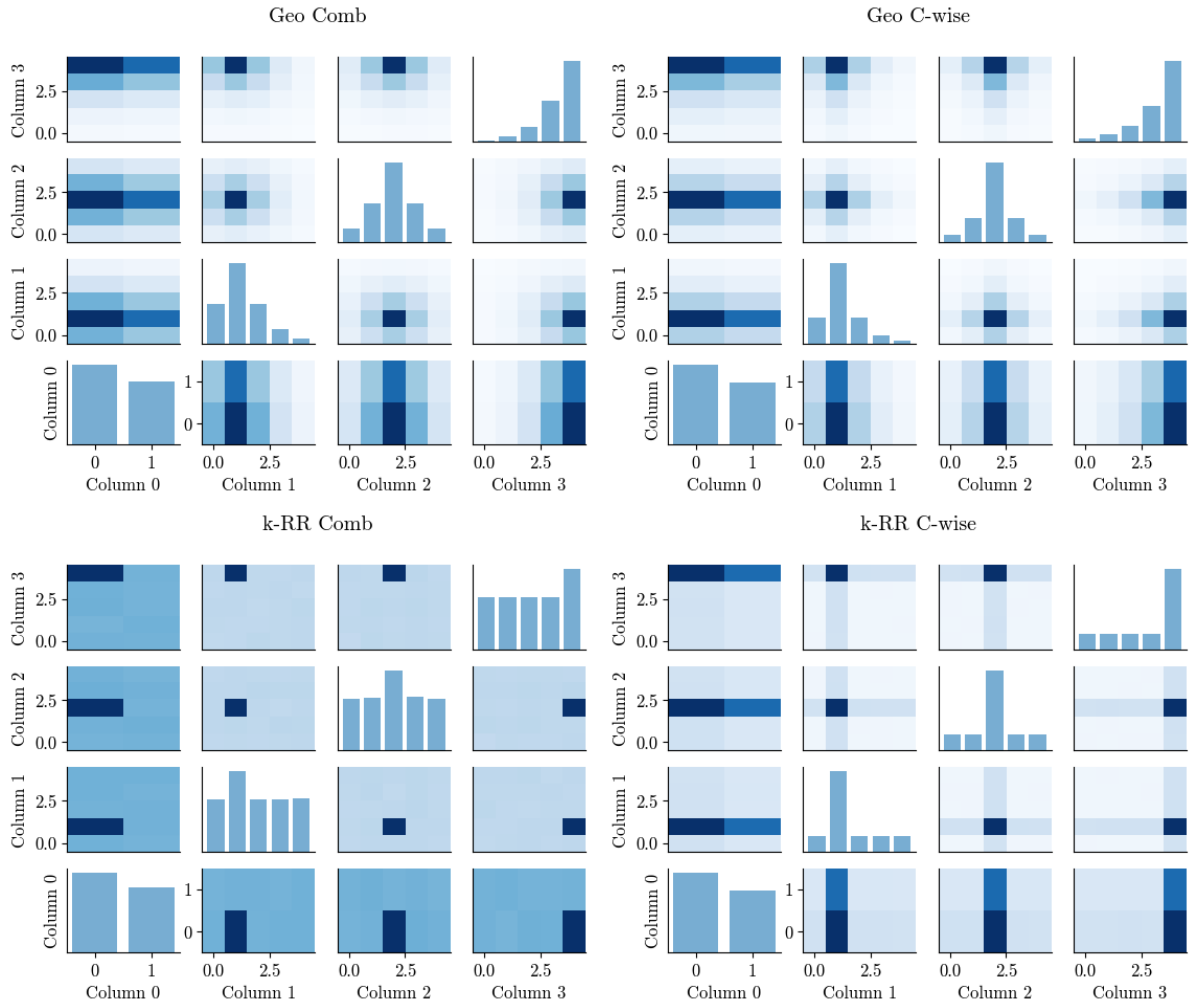


Figure C.1: Illustration of 4 multidimensional mechanisms discussed in this paper: 4D bounded Geometric, 4x1D bounded Geometric, 4D k-RR and 4x1D k-RR.

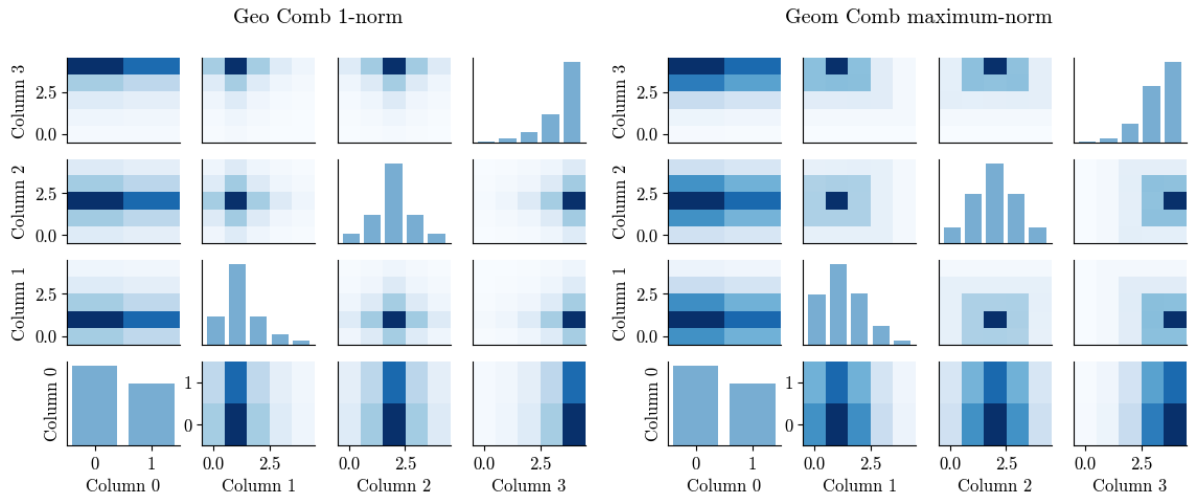


Figure C.2: Comparison between Manhattan ($p = 1$) and Chebyshev ($p = \infty$) distances for bounded geometric mechanisms. Refer to Figure C.1 for euclidean ($p = 2$).

C.2 Additional Experiments

We perform experiments using real and synthetic data. Data sets are distinguished into two main groups. The first category is pairwise data, which have two variables A and B where A causes B or B causes A. The task is to determine the causal direction between the two variables. The second category is the data that has more than two variables. The task here is to determine the causal structure (the skeleton) and the causal direction between the pairs within this structure.

C.2.1 F1 Score results Sachs data set

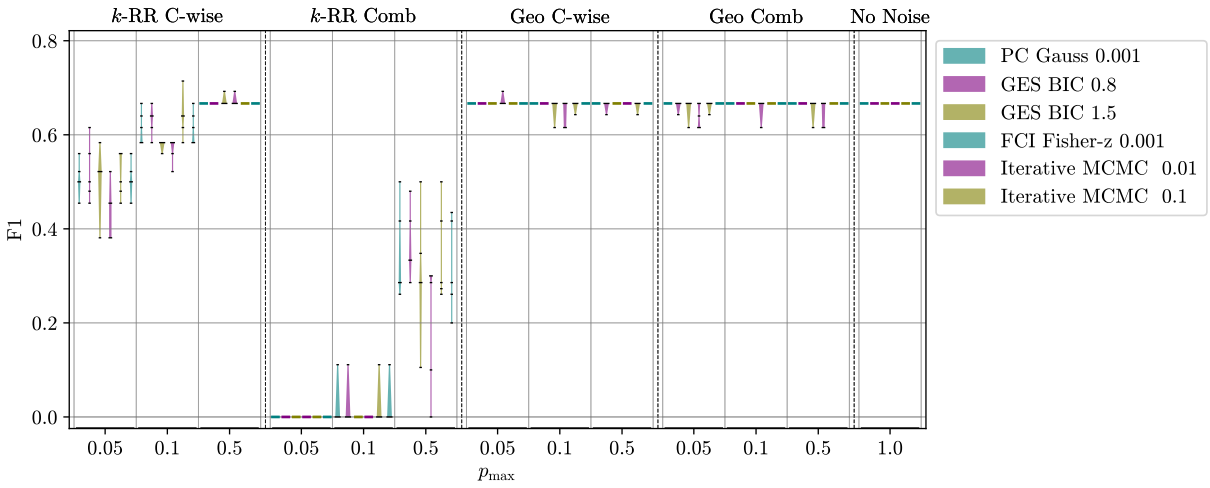
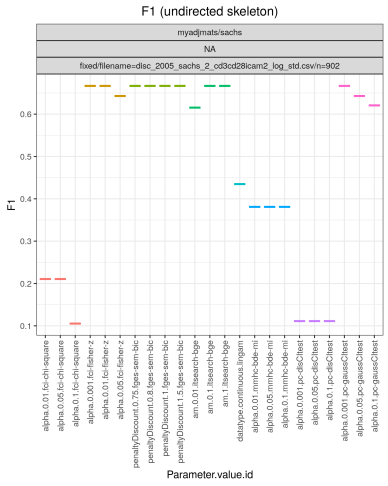
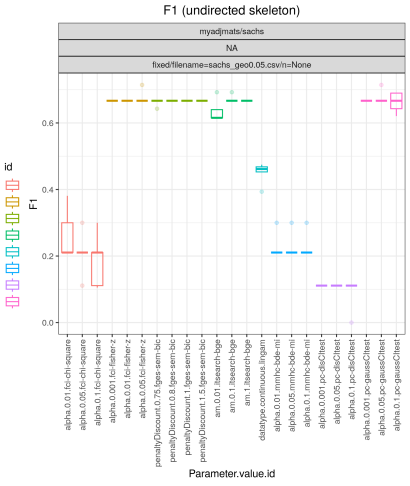


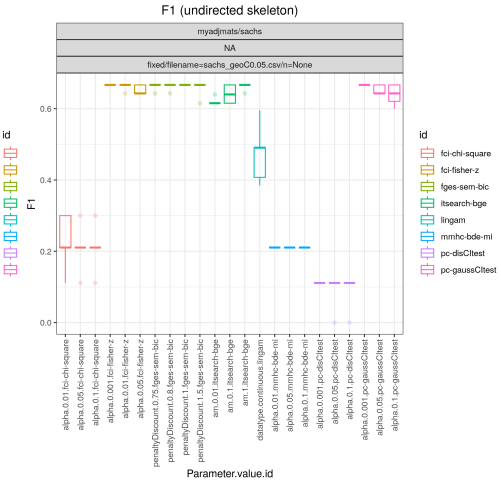
Figure C.3: Sachs data, all privacy methods, F1.



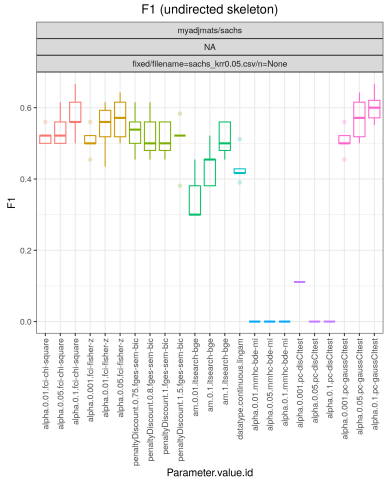
((a)) F1 Scores on the Sachs data set. Discretized, no noise.



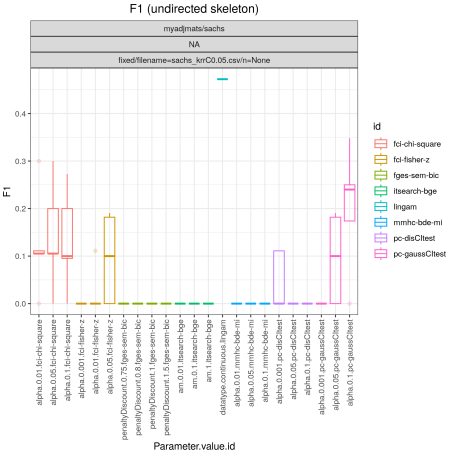
((b)) Sachs data, Geo C-wise mechanism, max probability 0.05.



((c)) Sachs data, Geo Comb mechanism, max probability 0.05.

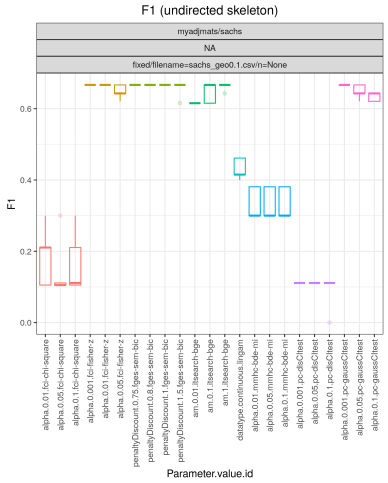


((d)) Sachs data, k-RR C-wise mechanism, max probability 0.05.

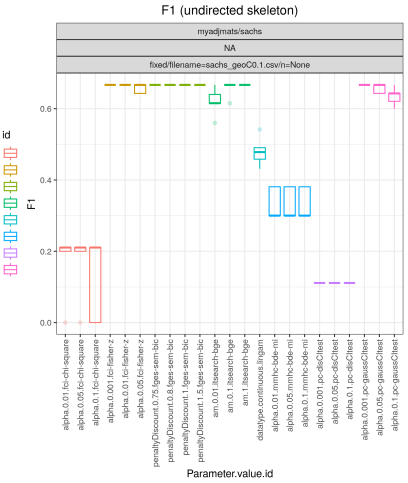


((e)) Sachs data, k-RR Comb mechanism, max probability 0.05.

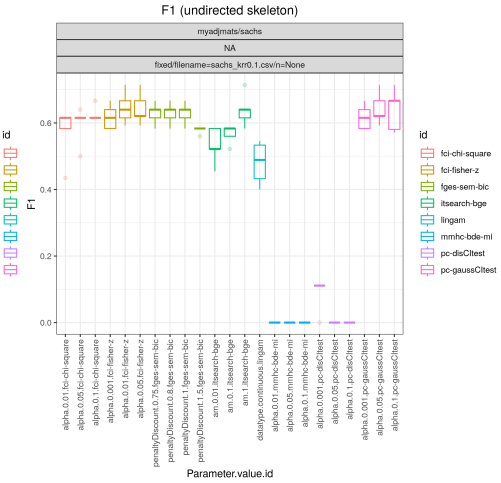
Figure C.4: Sachs data, F1, p-max 0.05.



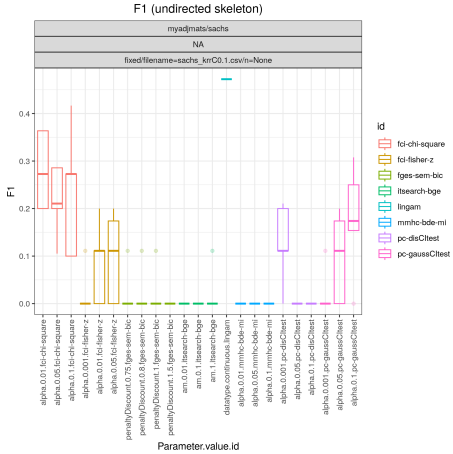
((a)) Sachs data, Geo C-wise mechanism, max probability 0.1.



((b)) Sachs data, Geo Comb mechanism, max probability 0.1.

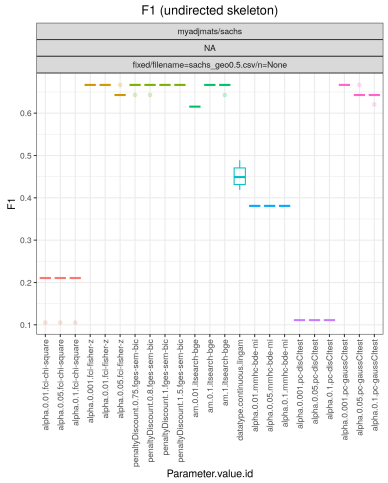


((c)) Sachs data, k-RR C-wise mechanism, max probability 0.1.

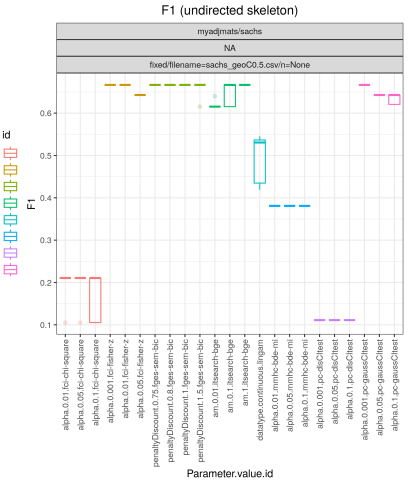


((d)) Sachs data, k-RR Comb mechanism, max probability 0.1.

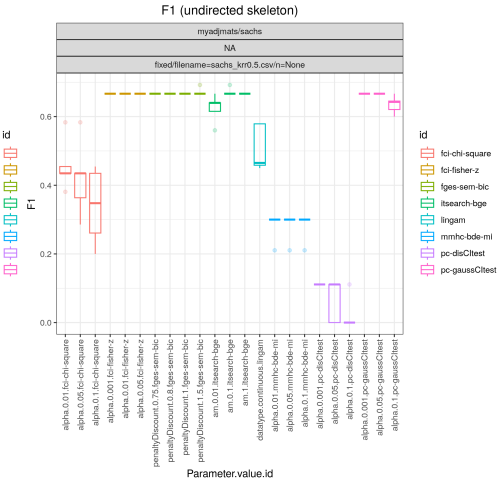
Figure C.5: Sachs data, F1, p-max 0.1



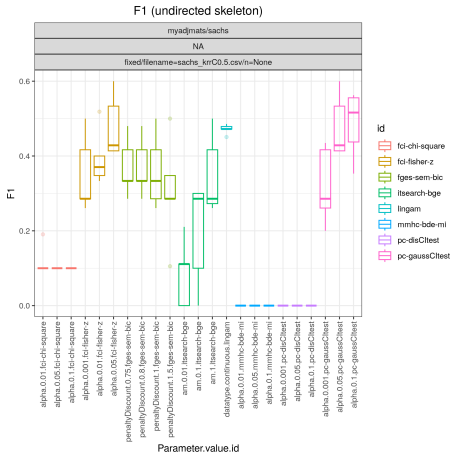
((a)) Sachs data, Geo C-wise mechanism, max probability 0.5.



((b)) Sachs data, Geo Comb mechanism, max probability 0.5.



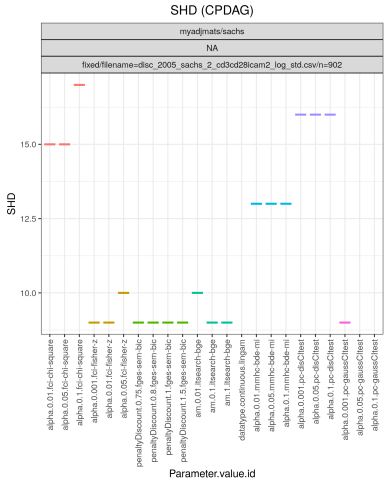
((c)) Sachs data, k-RR C-wise mechanism, max probability 0.5.



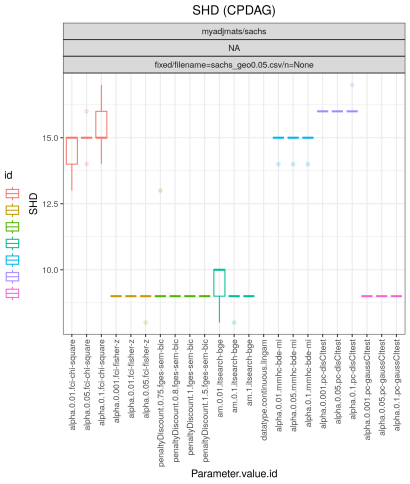
((d)) Sachs data, k-RR Comb mechanism, max probability 0.5.

Figure C.6: Sachs data, F1, p-max 0.5

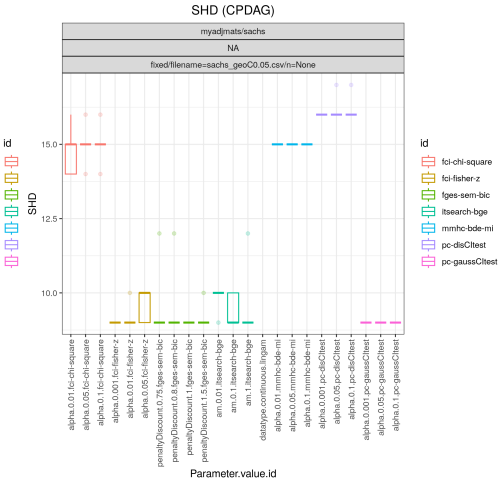
C.2.2 SHD Score results Sachs data set



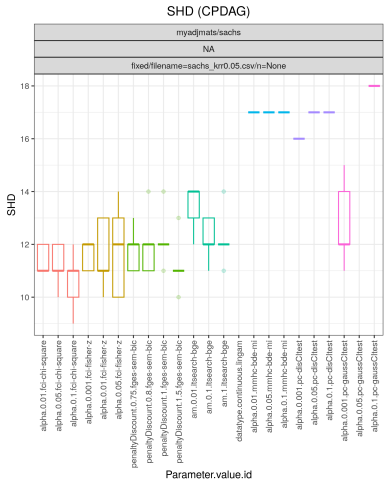
((a)) SHD Scores on the Sachs data set. Discretized, no noise.



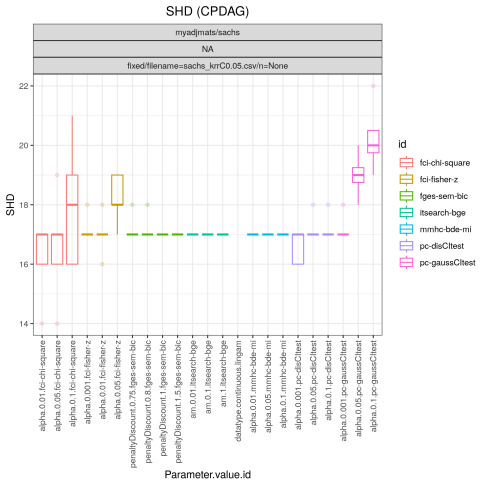
((b)) Sachs data, Geo C-wise mechanism, max probability 0.05.



((c)) Sachs data, Geo Comb mechanism, max probability 0.05.

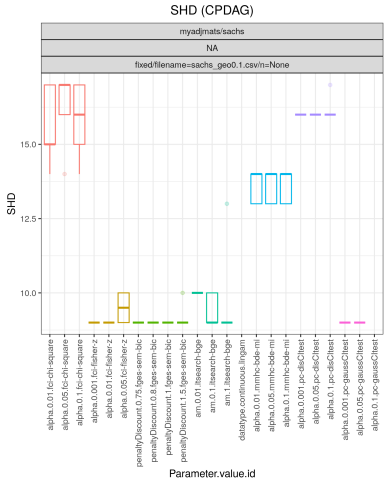


((d)) Sachs data, k-RR C-wise mechanism, max probability 0.05.

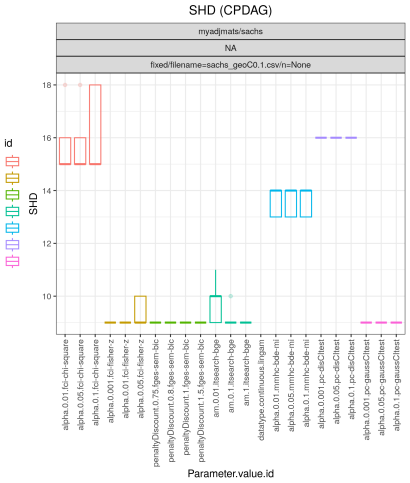


((e)) Sachs data, k-RR Comb mechanism, max probability 0.05.

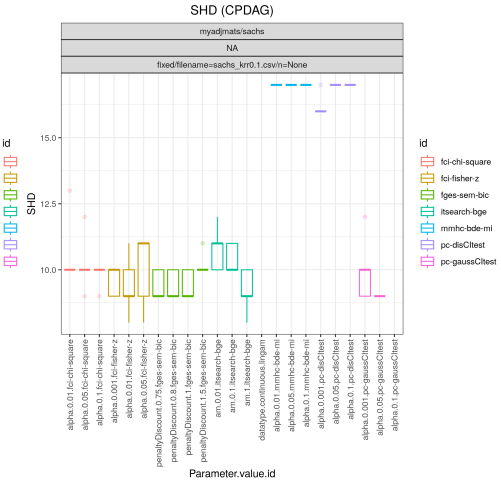
Figure C.7: Sachs data, SHD, p-max 0.05



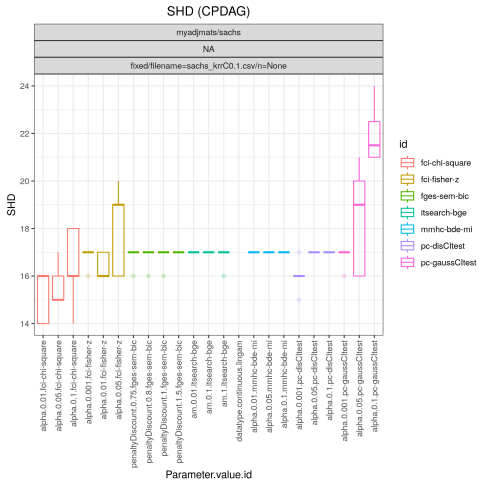
((a)) Sachs data, Geo C-wise mechanism, max probability 0.1.



((b)) Sachs data, Geo Comb mechanism, max probability 0.1.

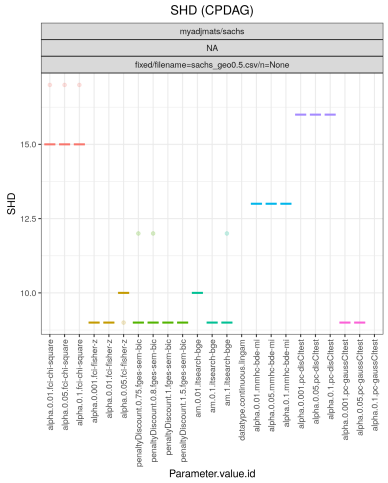


((c)) Sachs data, k-RR C-wise mechanism, max probability 0.1.

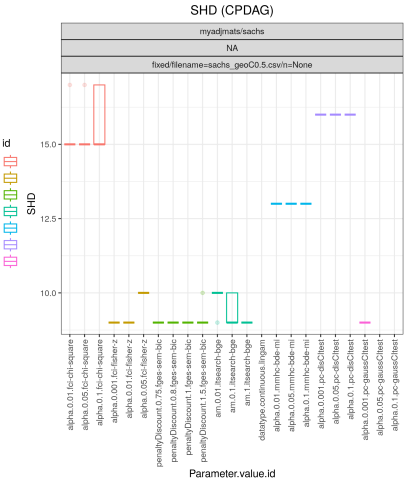


((d)) Sachs data, k-RR Comb mechanism, max probability 0.1.

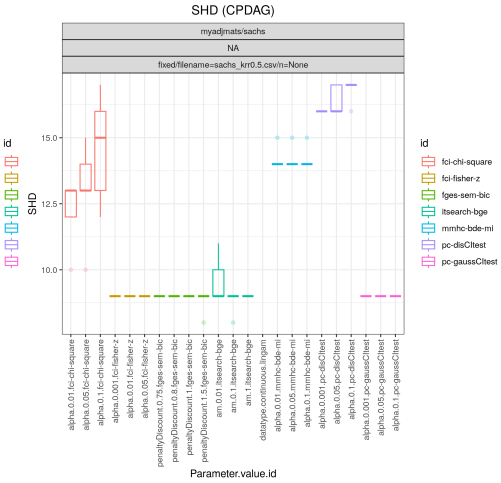
Figure C.8: Sachs data, SHD, p-max 0.1



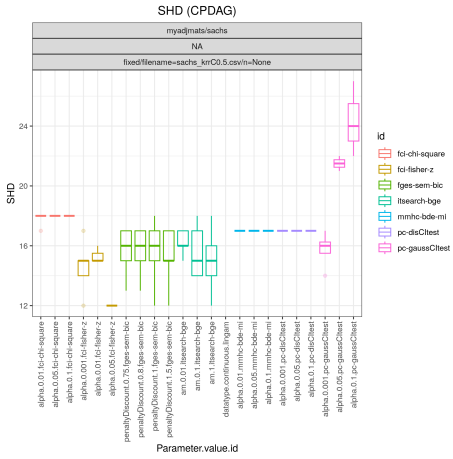
((a)) Sachs data, Geo C-wise mechanism, max probability 0.5.



((b)) Sachs data, Geo Comb mechanism, max probability 0.5.



((c)) Sachs data, k-RR C-wise mechanism, max probability 0.5.



((d)) Sachs data, k-RR Comb mechanism, max probability 0.5.

Figure C.9: Sachs data, SHD, p-max 0.5

C.2.3 F1 Score results Human Stature data set

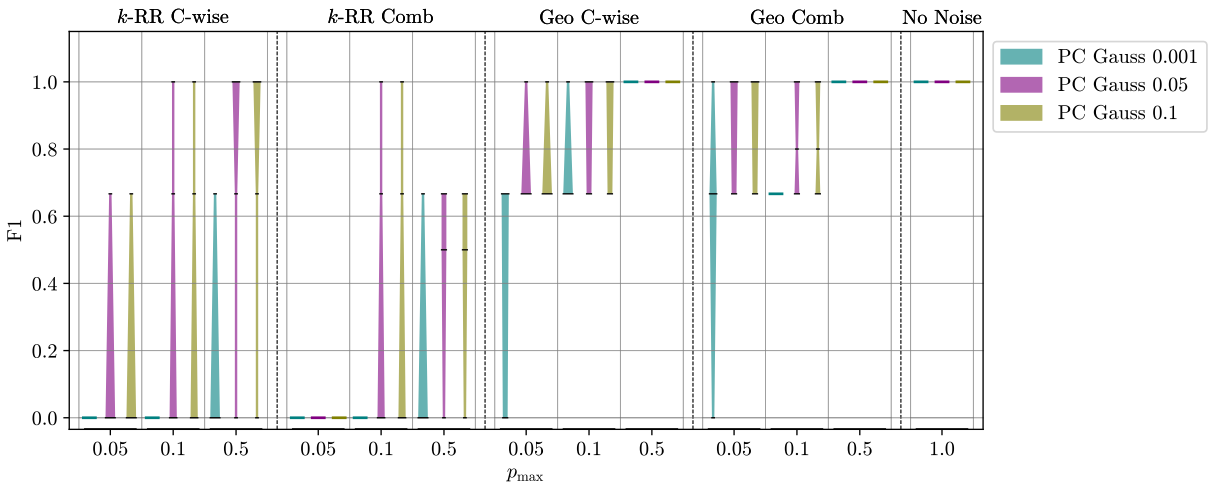
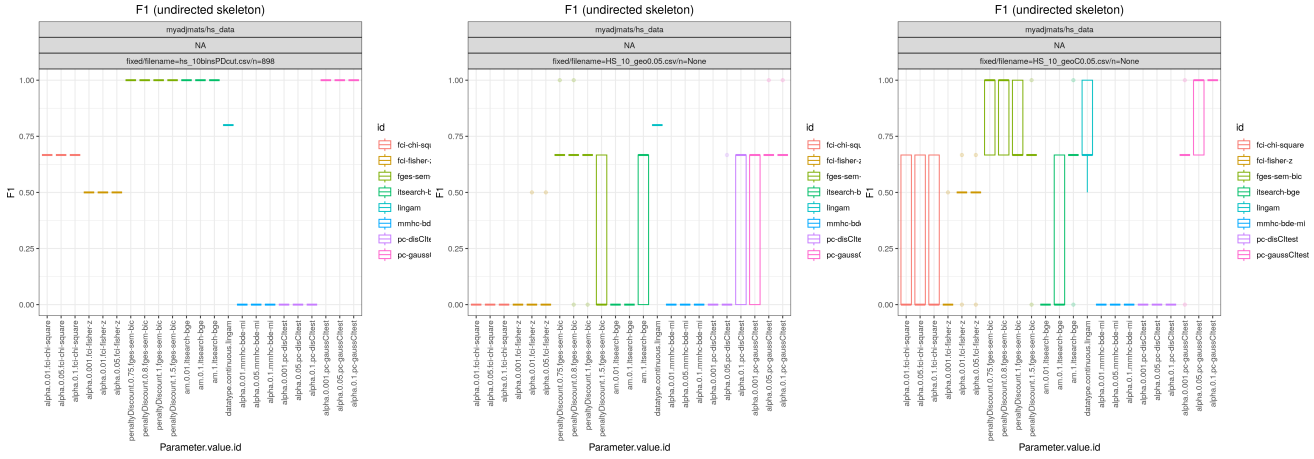


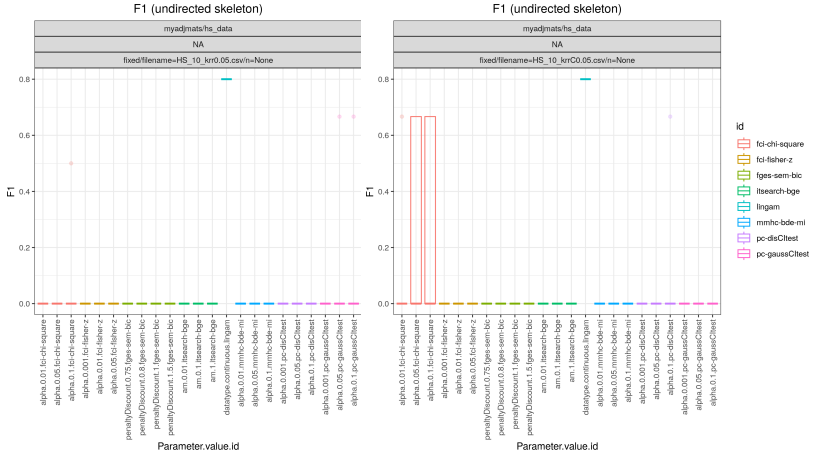
Figure C.10: Human Stature data, all privacy methods, F1.



((a)) F1 Scores on the Human Stature data set. Discretized, no noise.

((b)) Human Stature data, Geo C-wise mechanism, max probability 0.05.

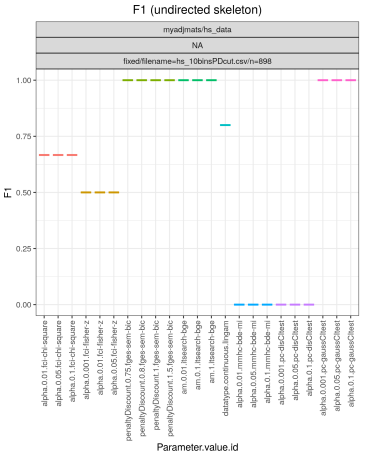
((c)) Human Stature data, Geo Comb mechanism, max probability 0.05.



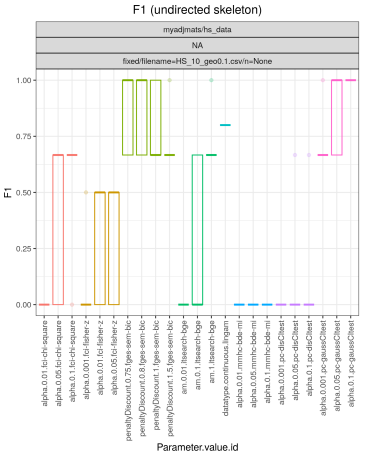
((d)) Human Stature data, k-RR C-wise mechanism, max probability 0.05.

((e)) Human Stature data, k-RR Comb mechanism, max probability 0.05.

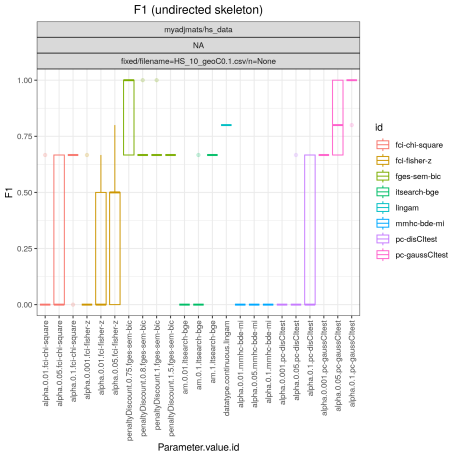
Figure C.11: Human Stature data, F1, p-max 0.05



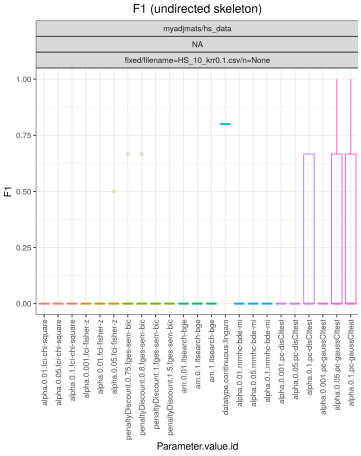
((a)) F1 Scores on the Human Stature data set. Discretized, no noise.



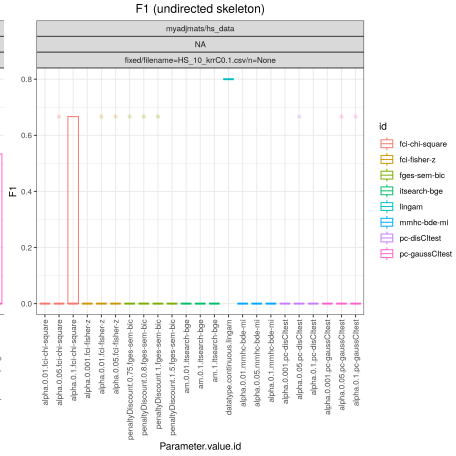
((b)) Human Stature data, Geo C-wise mechanism, max probability 0.1.



((c)) Human Stature data, Geo Comb mechanism, max probability 0.1.

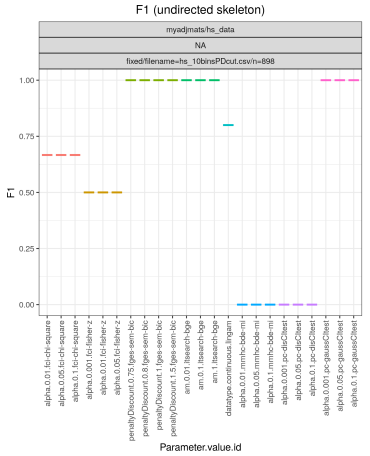


((d)) Human Stature data, k-RR C-wise mechanism, max probability 0.1.

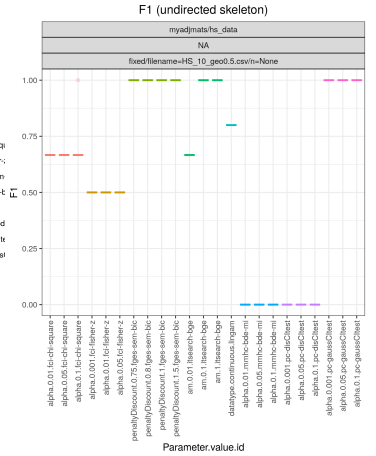


((e)) Human Stature data, k-RR Comb mechanism, max probability 0.1.

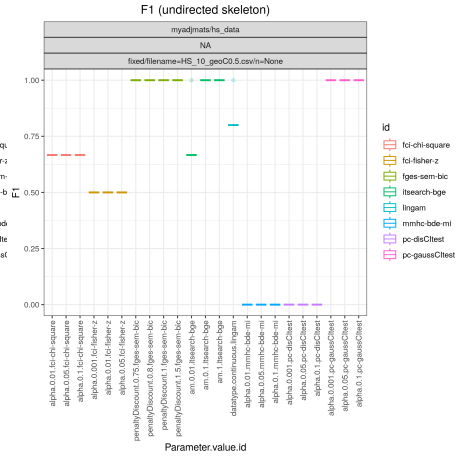
Figure C.12: Human Stature data, F1, p-max 0.1



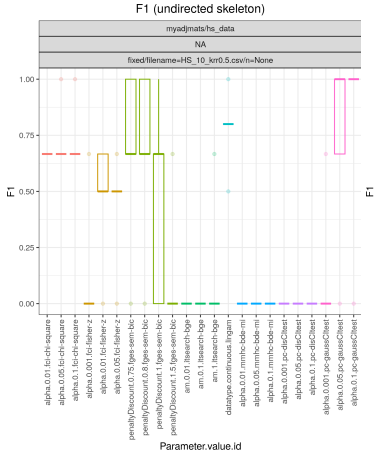
((a)) F1 Scores on the Human Stature data set. Discretized, no noise.



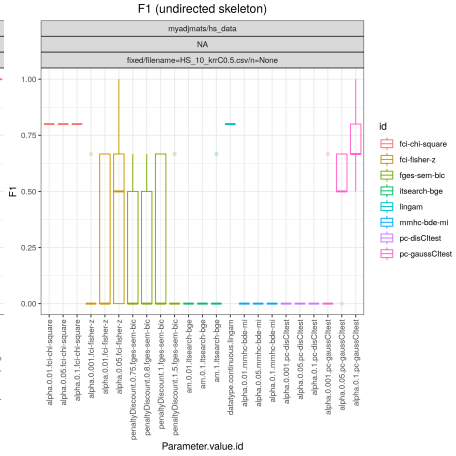
((b)) Human Stature data, Geo C-wise mechanism, max probability 0.5.



((c)) Human Stature data, Geo Comb mechanism, max probability 0.5.



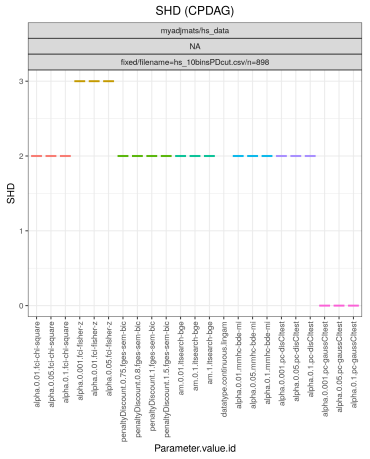
((d)) Human Stature data, k-RR C-wise mechanism, max probability 0.5.



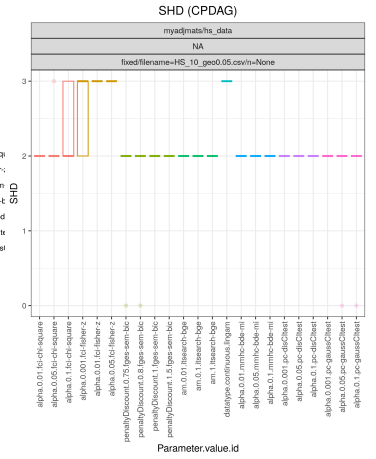
((e)) Human Stature data, k-RR Comb mechanism, max probability 0.5.

Figure C.13: Human Stature data, F1, p-max 0.5

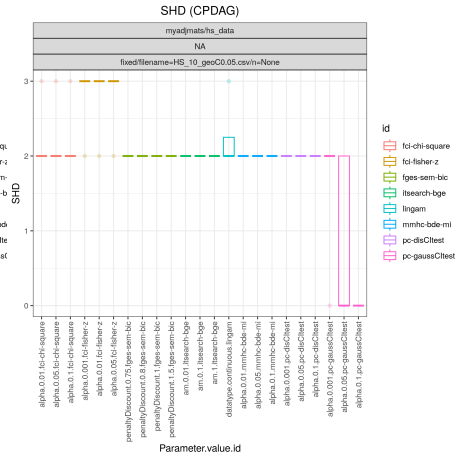
C.2.4 SHD Score results Human Stature data set



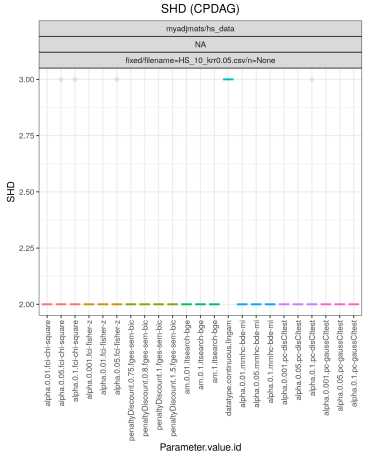
((a)) SHD Scores on the Human Stature data set. Discretized, no noise.



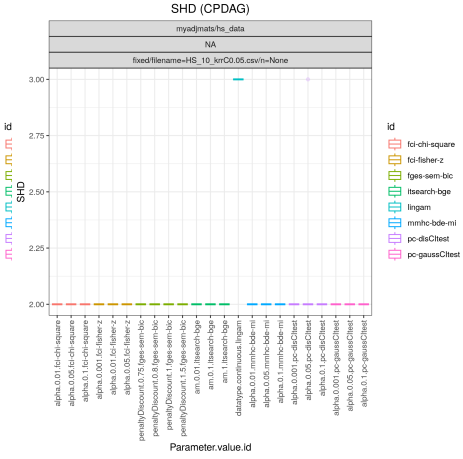
((b)) Human Stature data, Geo C-wise mechanism, max probability 0.05.



((c)) Human Stature data, Geo Comb mechanism, max probability 0.05.

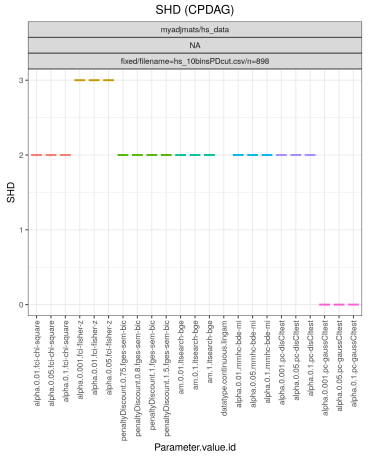


((d)) Human Stature data, k-RR C-wise mechanism, max probability 0.05.

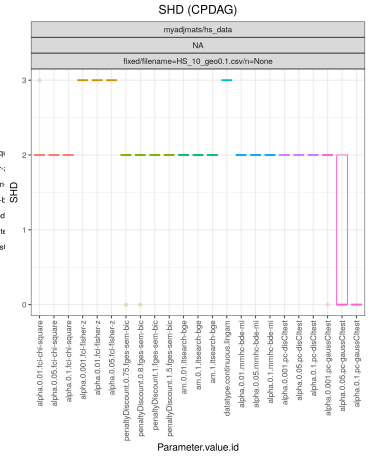


((e)) Human Stature data, k-RR Comb mechanism, max probability 0.05.

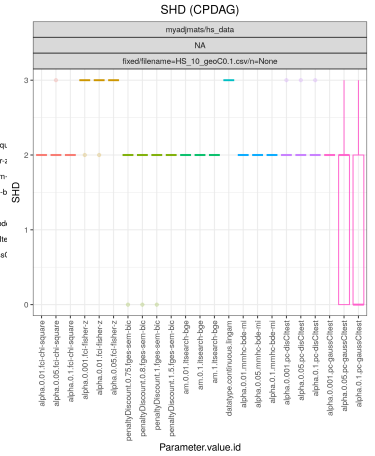
Figure C.14: Human Stature data, SHD, p-max 0.05



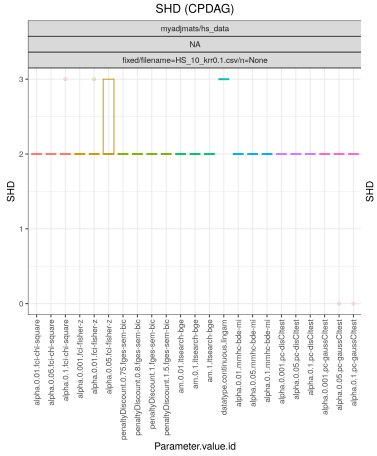
((a)) F1 Scores on the Human Stature data set. Discretized, no noise.



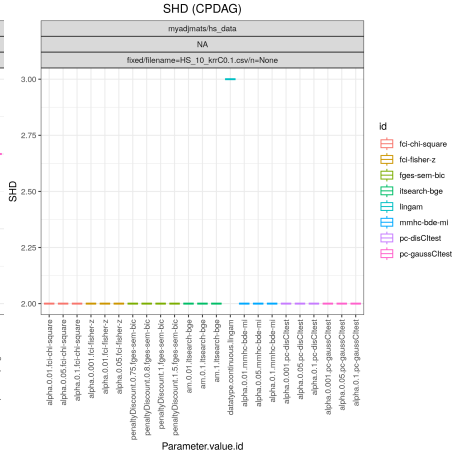
((b)) Human Stature data, Geo C-wise mechanism, max probability 0.1.



((c)) Human Stature data, Geo Comb mechanism, max probability 0.1.

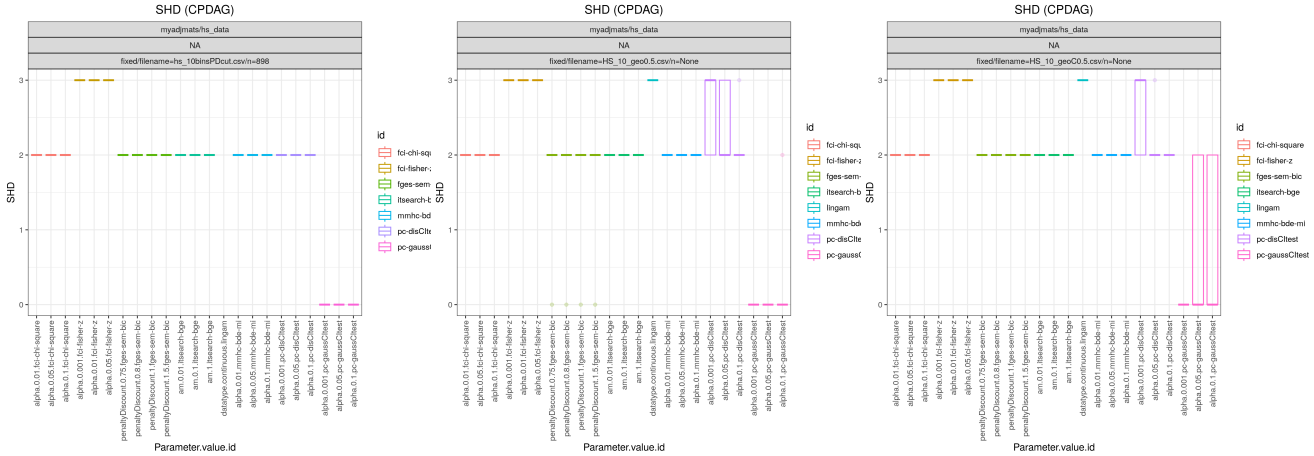


((d)) Human Stature data, k-RR C-wise mechanism, max probability 0.1.



((e)) Human Stature data, k-RR Comb mechanism, max probability 0.1.

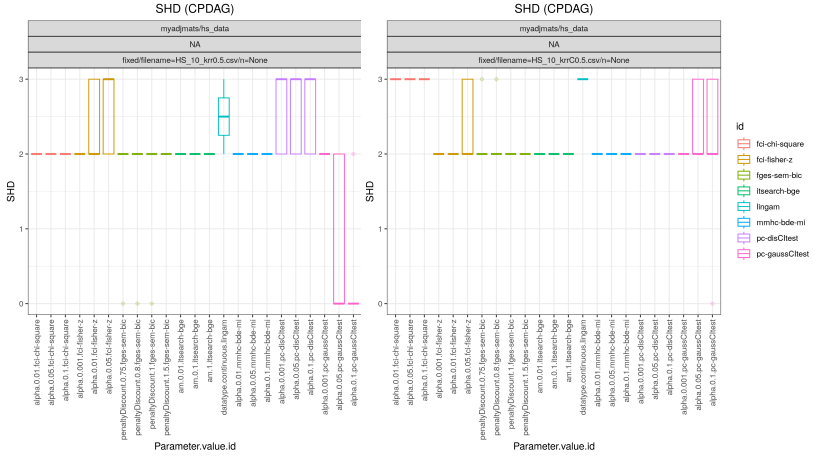
Figure C.15: Human Stature data, SHD, p-max 0.1



((a)) SHD Scores on the Human Stature data set. Discretized, no noise.

((b)) Human Stature data, Geo C-wise mechanism, max probability 0.5.

((c)) Human Stature data, Geo Comb mechanism, max probability 0.5.



((d)) Human Stature data, k-RR C-wise mechanism, max probability 0.5.

((e)) Human Stature data, k-RR Comb mechanism, max probability 0.5.

Figure C.16: Human Stature data, SHD, p-max 0.5

C.2.5 F1 Score results Synthetic 5 nodes data set

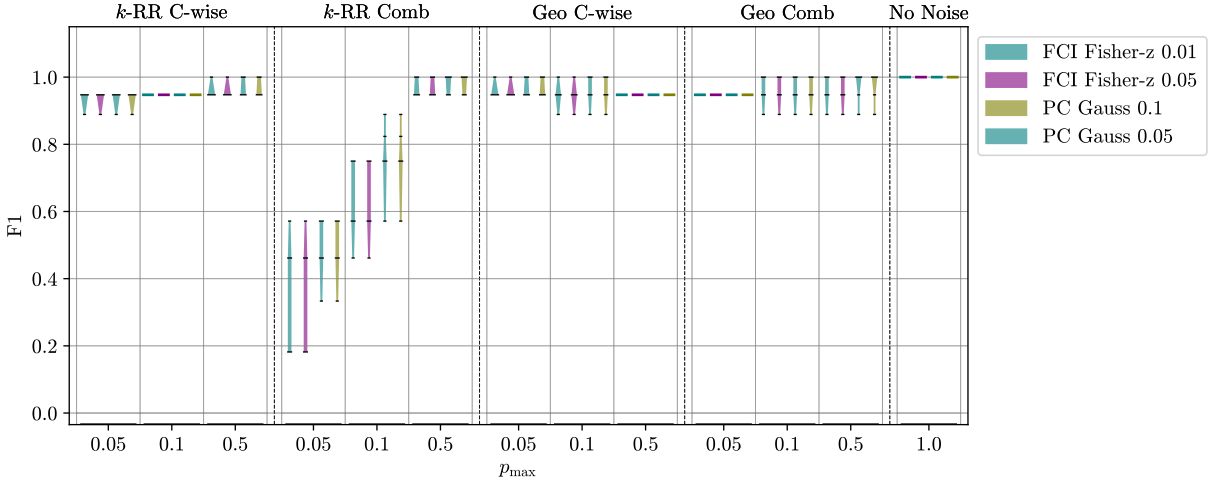
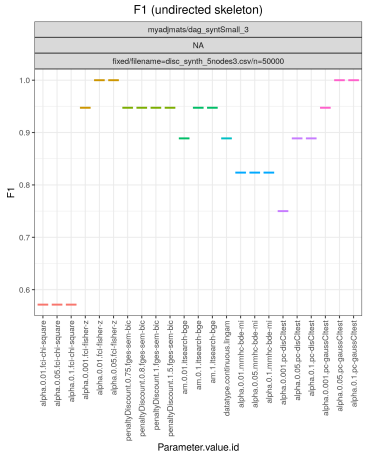
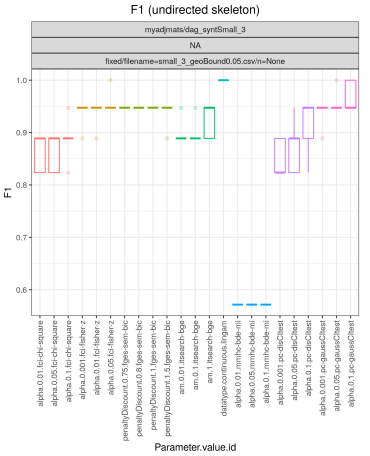


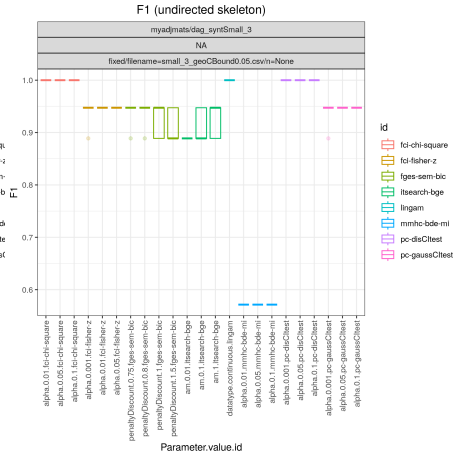
Figure C.17: Synthetic data, 5 nodes, all privacy methods, F1.



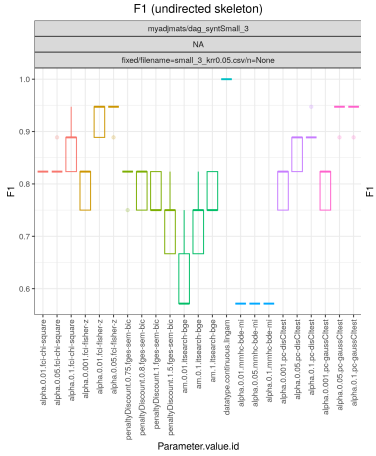
((a)) F1 Scores on the Synthetic 5 nodes data set. Discretized, no noise.



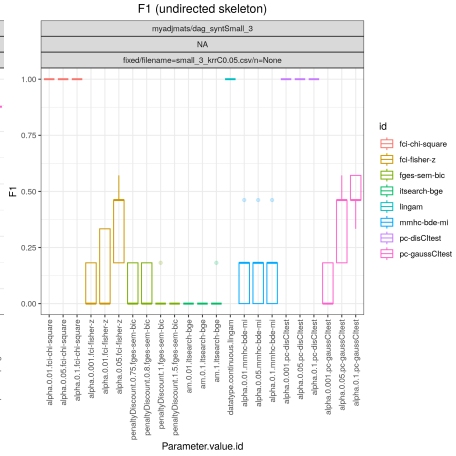
((b)) Synthetic 5 nodes data, Geo C-wise mechanism, max probability 0.05.



((c)) Synthetic 5 nodes data, Geo Comb mechanism, max probability 0.05.

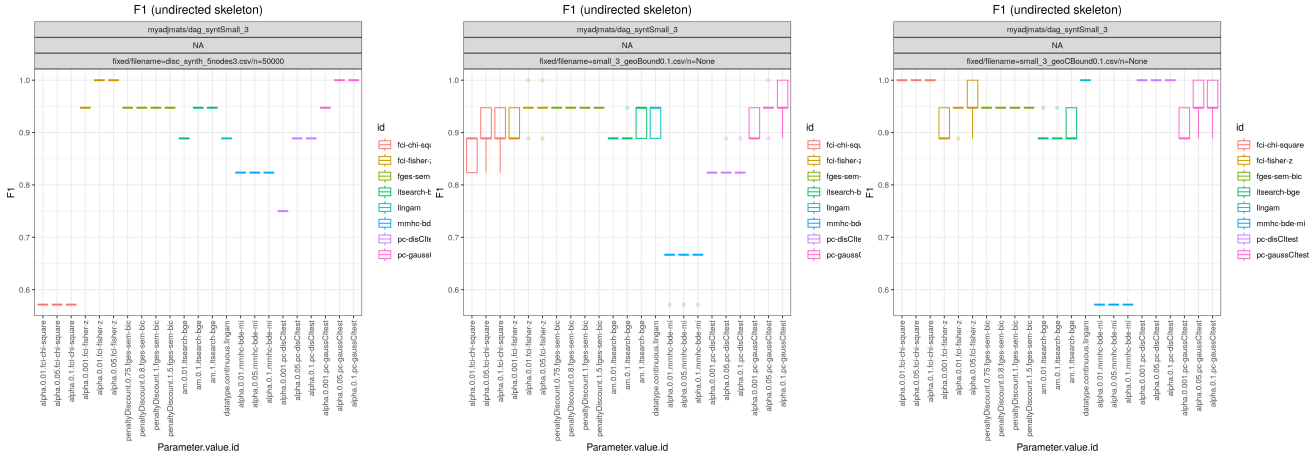


((d)) Synthetic 5 nodes data, k-RR C-wise mechanism, max probability 0.05.



((e)) Synthetic 5 nodes data, k-RR Comb mechanism, max probability 0.05.

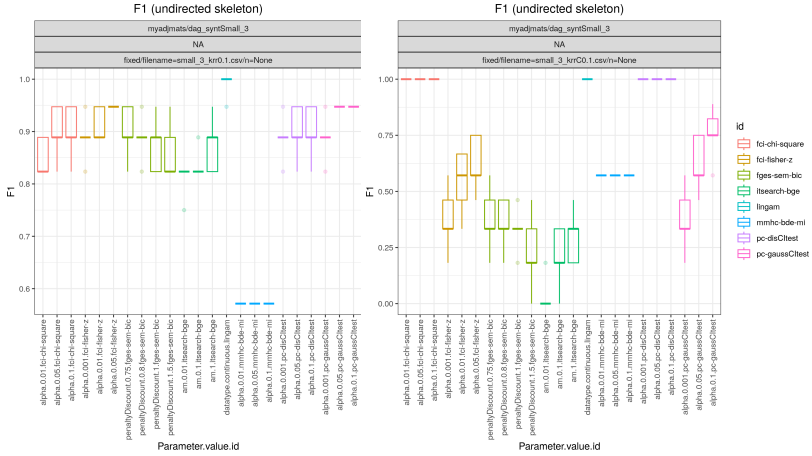
Figure C.18: Synth5 data, F1, p-max 0.05



((a)) F1 Scores on the Synthetic 5 nodes data set. Discretized, no noise.

((b)) Synthetic 5 nodes data, Geo C-wise mechanism, max probability 0.1.

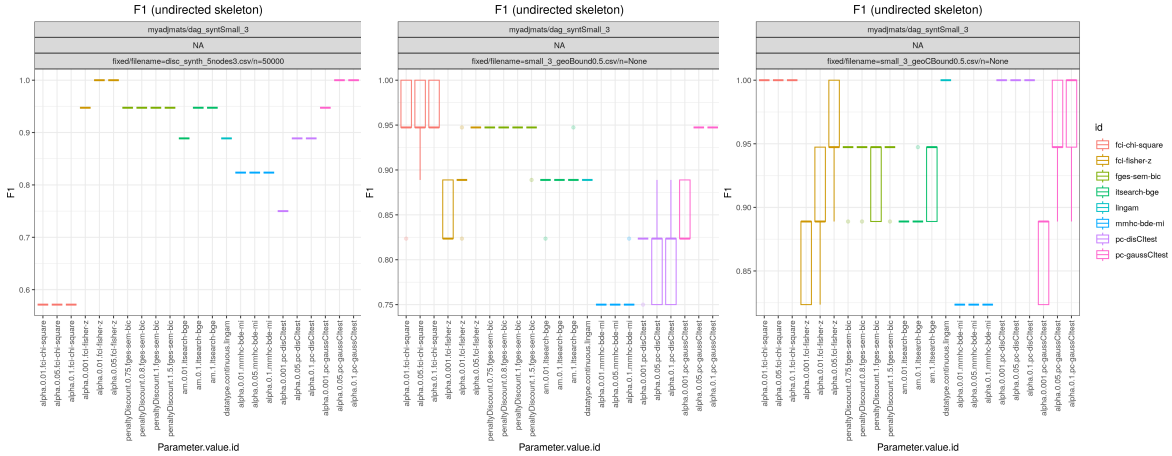
((c)) Synthetic 5 nodes data, Geo Comb mechanism, max probability 0.1.



((d)) Synthetic 5 nodes data, k-RR C-wise mechanism, max probability 0.1.

((e)) Synthetic 5 nodes data, k-RR Comb mechanism, max probability 0.1.

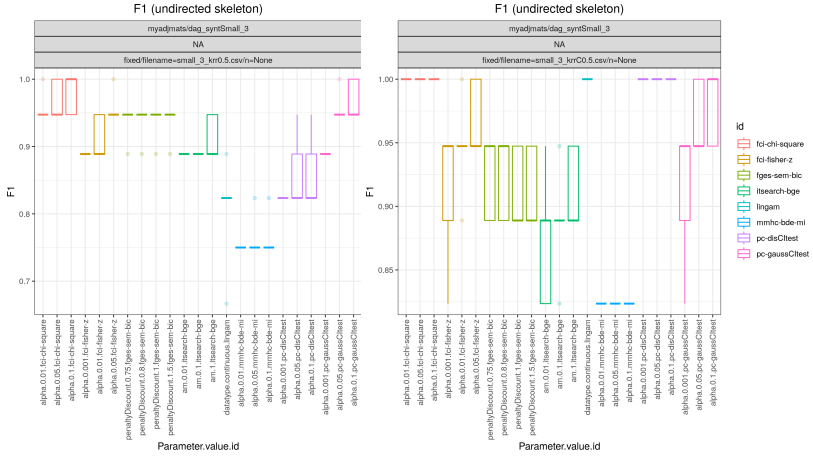
Figure C.19: Synth5 data, F1, p-max 0.1



((a)) F1 Scores on the Synthetic 5 nodes data set. Discretized, no noise.

((b)) Synthetic 5 nodes data, Geo C-wise mechanism, max probability 0.5.

((c)) Synthetic 5 nodes data, Geo Comb mechanism, max probability 0.5.

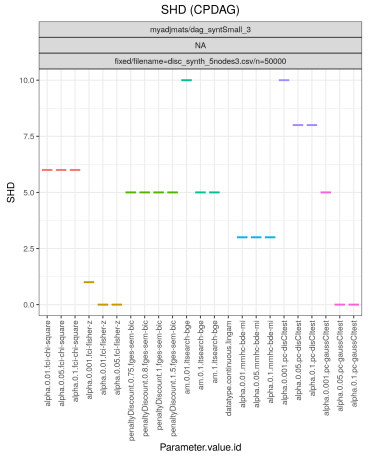


((d)) Synthetic 5 nodes data, k-RR C-wise mechanism, max probability 0.5.

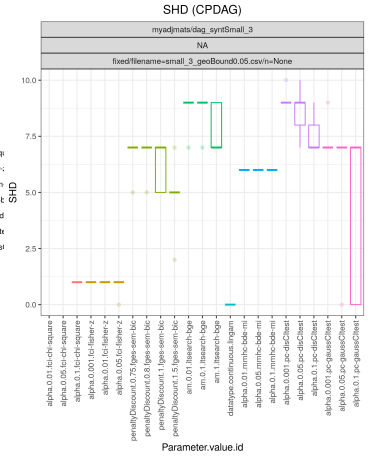
((e)) Synthetic 5 nodes data, k-RR Comb mechanism, max probability 0.5.

Figure C.20: Synth5 data, F1, p-max 0.5

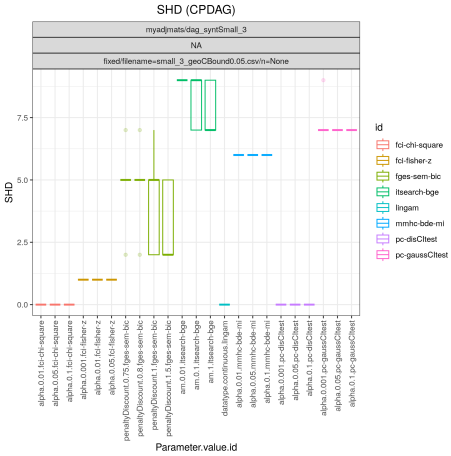
C.2.6 SHD Score results Synthetic 5 nodes data set



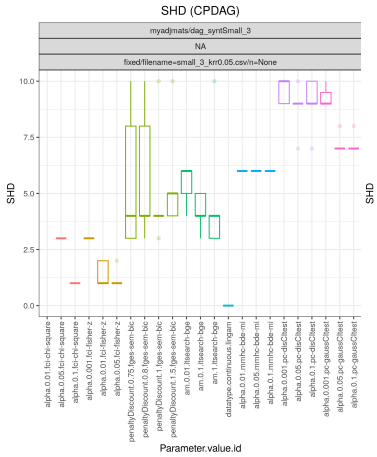
((a)) SHD Scores on the Synthetic 5 nodes data set. Discretized, no noise.



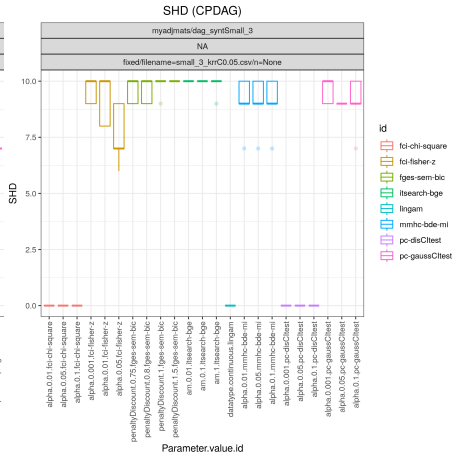
((b)) Synthetic 5 nodes data, Geo C-wise mechanism, max probability 0.05.



((c)) Synthetic 5 nodes data, Geo Comb mechanism, max probability 0.05.

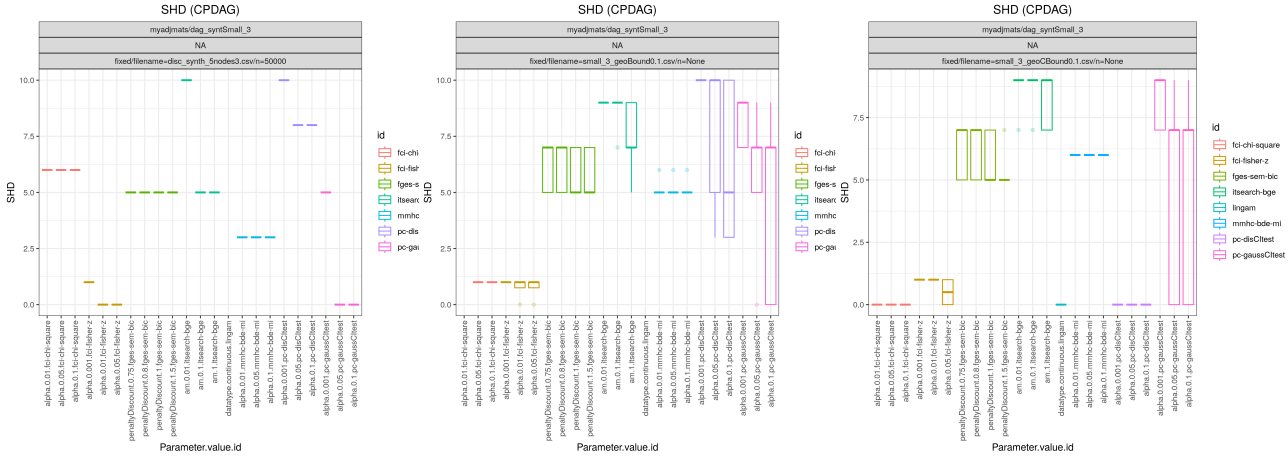


((d)) Synthetic 5 nodes data, k-RR C-wise mechanism, max probability 0.05.



((e)) Synthetic 5 nodes data, k-RR Comb mechanism, max probability 0.05.

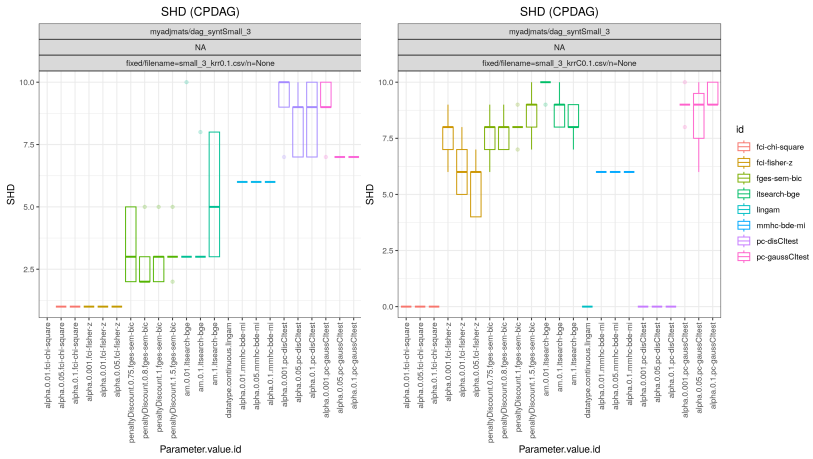
Figure C.21: Synth5 data, SHD, p-max 0.05



((a)) SHD Scores on the Synthetic 5 nodes data set. Discretized, no noise.

((b)) Synthetic 5 nodes data, Geo C-wise mechanism, max probability 0.1.

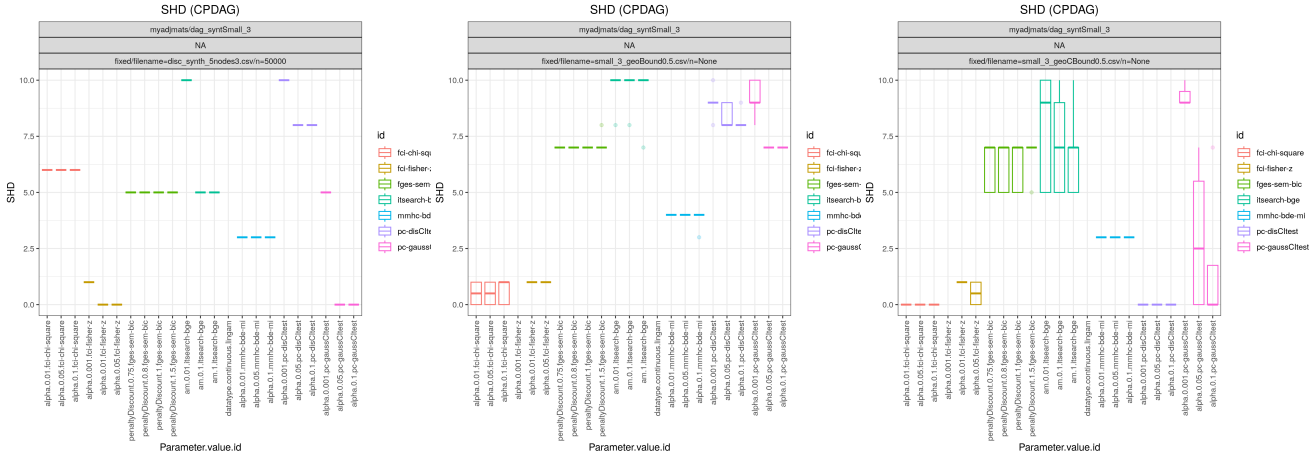
((c)) Synthetic 5 nodes data, Geo Comb mechanism, max probability 0.1.



((d)) Synthetic 5 nodes data, k-RR C-wise mechanism, max probability 0.1.

((e)) Synthetic 5 nodes data, k-RR Comb mechanism, max probability 0.1.

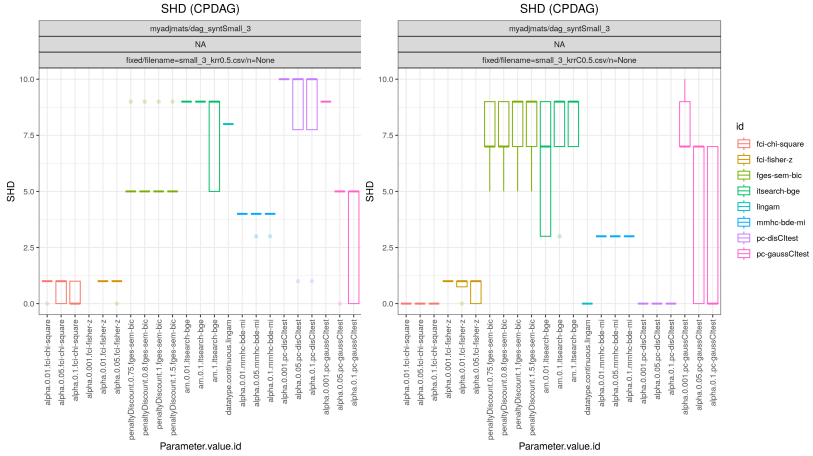
Figure C.22: Synth5 data, SHD, p-max 0.1



((a)) SHD Scores on the Synthetic 5 nodes data set. Discretized, no noise.

((b)) Synthetic 5 nodes data, Geo C-wise mechanism, max probability 0.5.

((c)) Synthetic 5 nodes data, Geo Comb mechanism, max probability 0.5.



((d)) Synthetic 5 nodes data, k-k-RR C-wise mechanism, max probability 0.5.

((e)) Synthetic 5 nodes data, k-RR Comb mechanism, max probability 0.5.

Figure C.23: Synth5 data, SHD, p-max 0.5

C.2.7 F1 Score results Synthetic 10 nodes data set

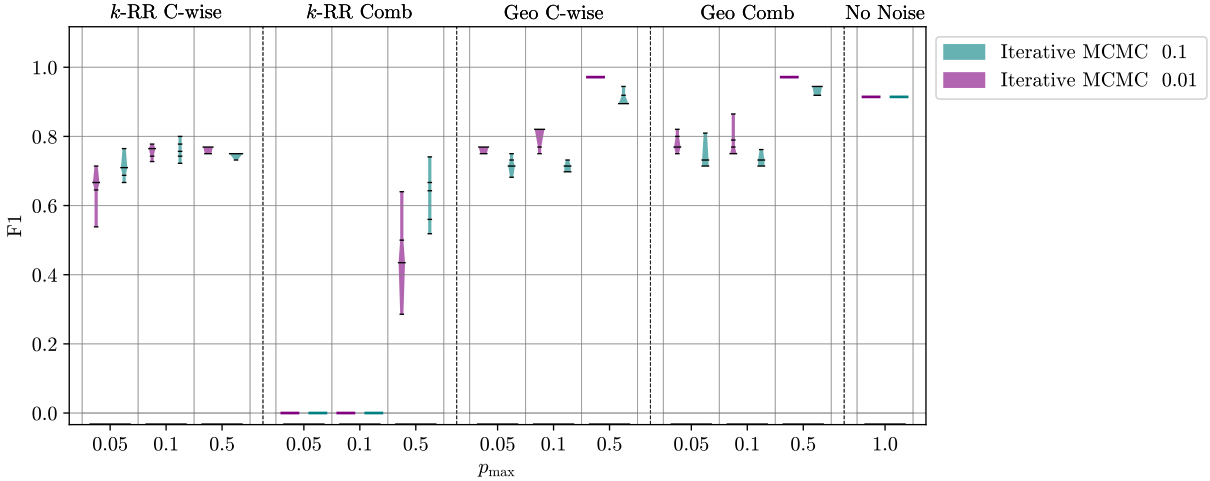
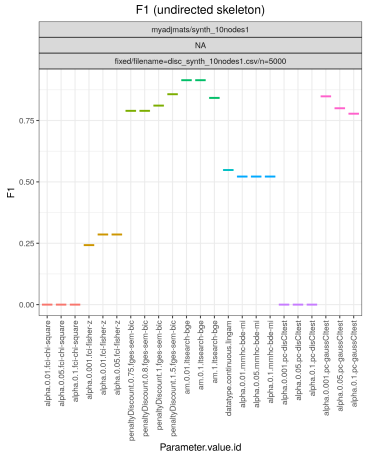
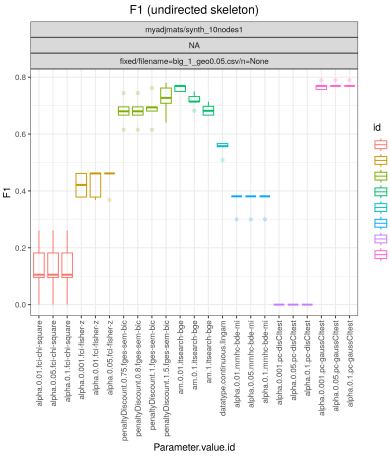


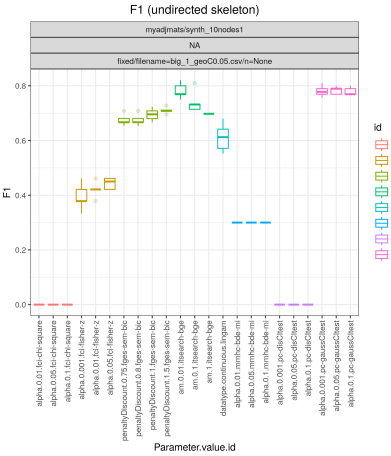
Figure C.24: Synthetic data, 10 nodes, all privacy methods, F1.



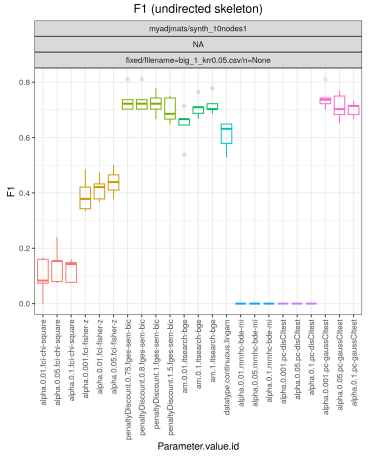
((a)) F1 Scores on the Synthetic 10 nodes data set. Discretized, no noise.



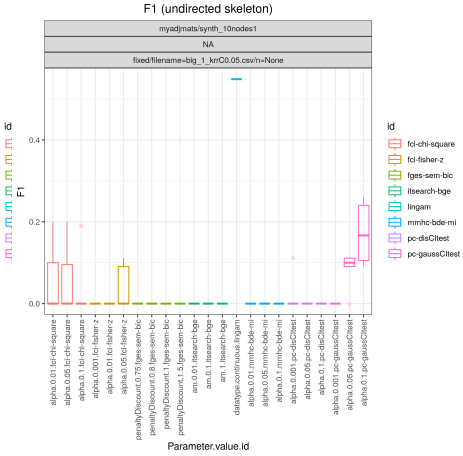
((b)) Synthetic 10 nodes data, Geo C-wise mechanism, max probability 0.05.



((c)) Synthetic 10 nodes data, Geo Comb mechanism, max probability 0.05.

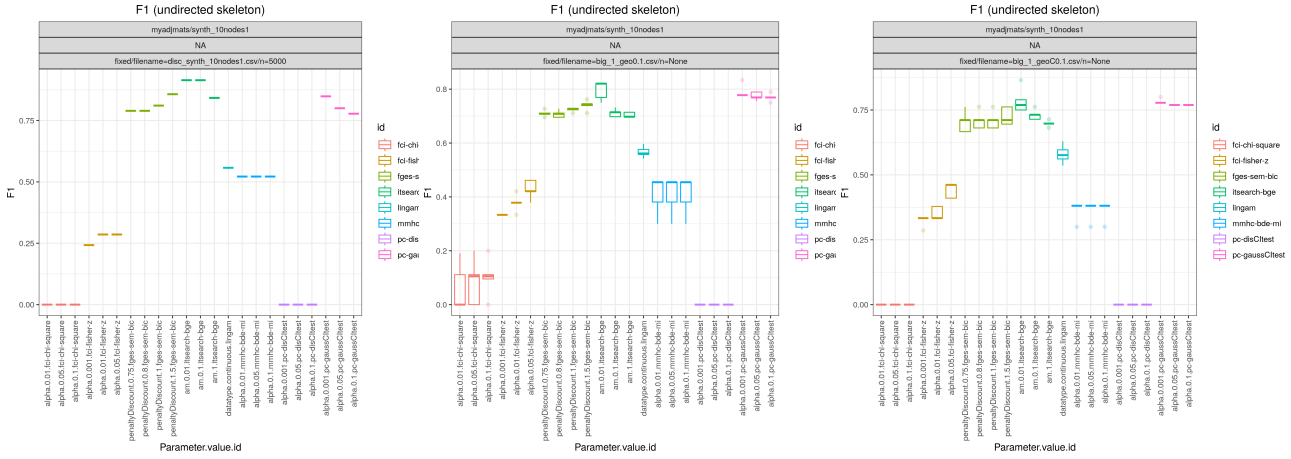


((d)) Synthetic 10 nodes data, k-RR C-wise mechanism, max probability 0.05.



((e)) Synthetic 10 nodes data, k-RR Comb mechanism, max probability 0.05.

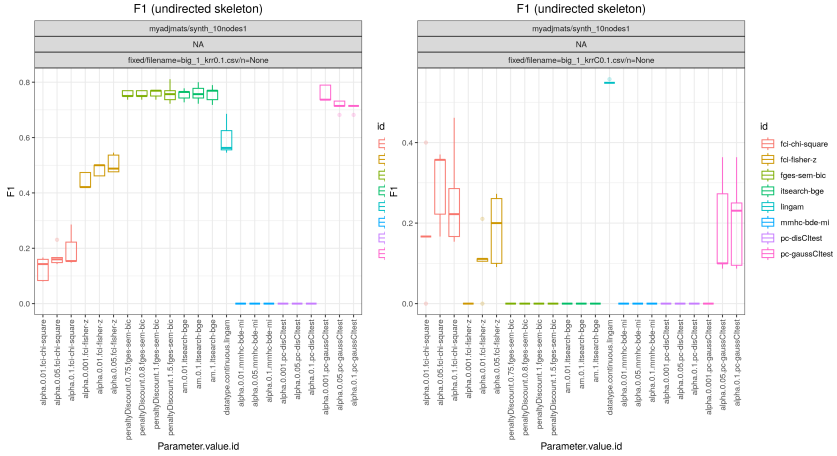
Figure C.25: Synth10 data, F1, p-max 0.05



((a)) F1 Scores on the Synthetic 10 nodes data set. Discretized, no noise.

((b)) Synthetic 10 nodes data, Geo C-wise mechanism, max probability 0.1.

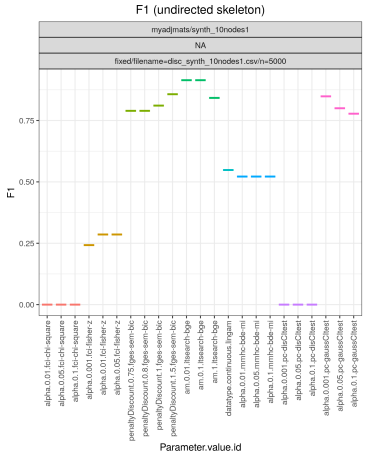
((c)) Synthetic 10 nodes data, Geo Comb mechanism, max probability 0.1.



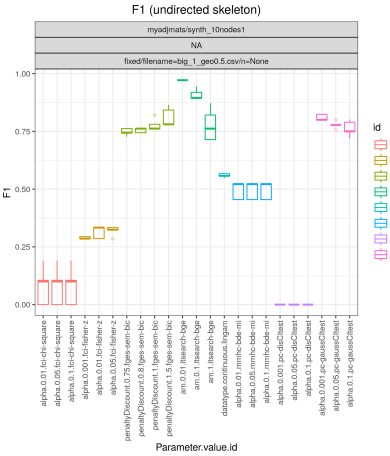
((d)) Synthetic 10 nodes data, k-RR C-wise mechanism, max probability 0.1.

((e)) Synthetic 10 nodes data, k-RR Comb mechanism, max probability 0.1.

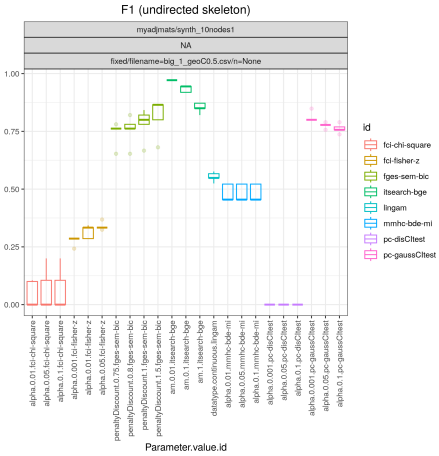
Figure C.26: Synth10 data, F1, p-max 0.1



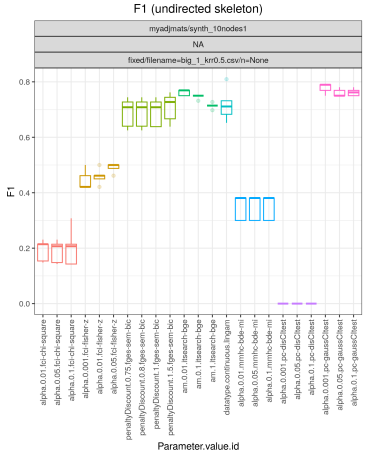
((a)) F1 Scores on the Synthetic 10 nodes data set. Discretized, no noise.



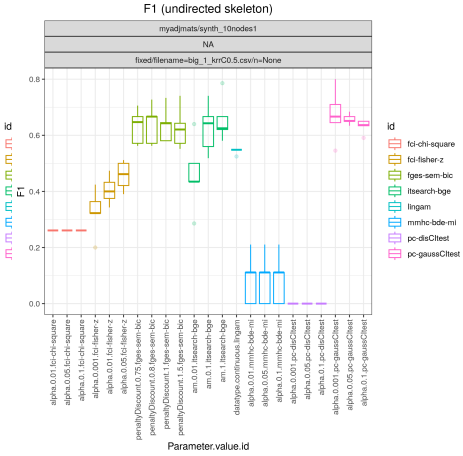
((b)) Synthetic 10 nodes data, Geo C-wise mechanism, max probability 0.5.



((c)) Synthetic 10 nodes data, Geo Comb mechanism, max probability 0.5.



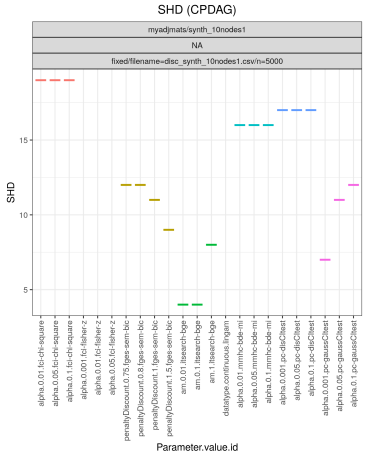
((d)) Synthetic 10 nodes data, k-RR C-wise mechanism, max probability 0.5.



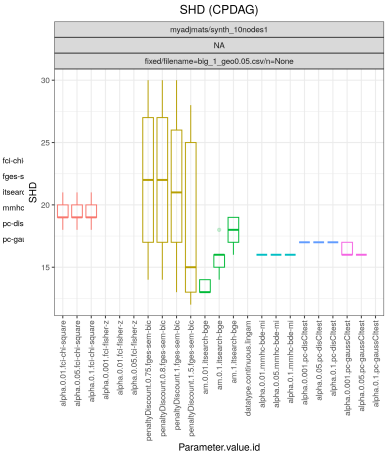
((e)) Synthetic 10 nodes data, k-RR Comb mechanism, max probability 0.5.

Figure C.27: Synth10 data, F1, p-max 0.5

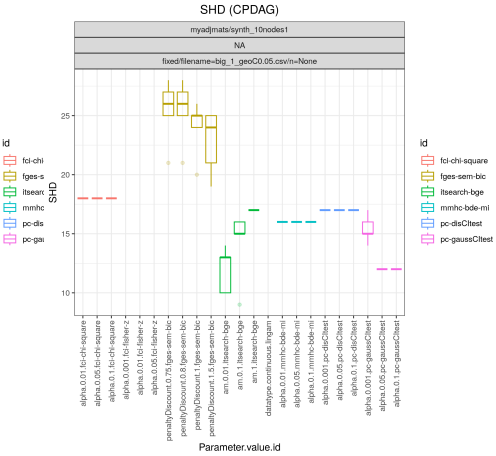
C.2.8 SHD Score results Synthetic 10 nodes data set



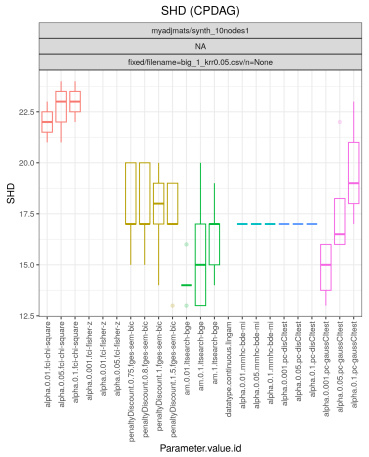
((a)) SHD Scores on the Synthetic 10 nodes data set. Discretized, no noise.



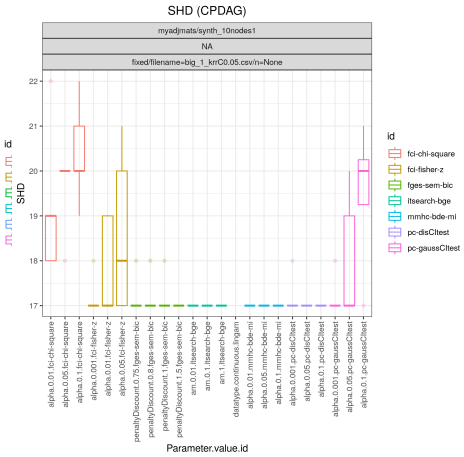
((b)) Synthetic 10 nodes data, Geo C-wise mechanism, max probability 0.05.



((c)) Synthetic 10 nodes data, Geo Comb mechanism, max probability 0.05.

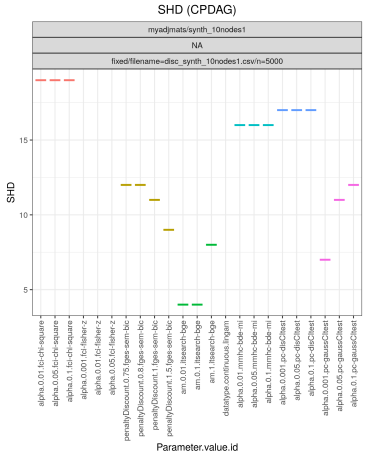


((d)) Synthetic 10 nodes data, k-RR C-wise mechanism, max probability 0.05.

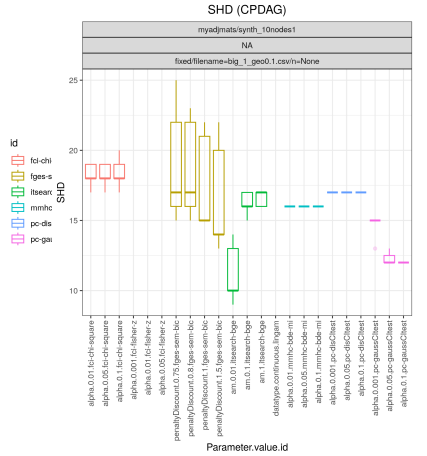


((e)) Synthetic 10 nodes data, k-RR Comb mechanism, max probability 0.05.

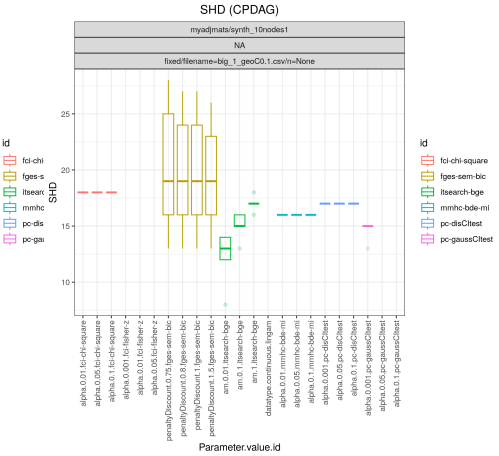
Figure C.28: Synth10 data, SHD, p-max 0.05



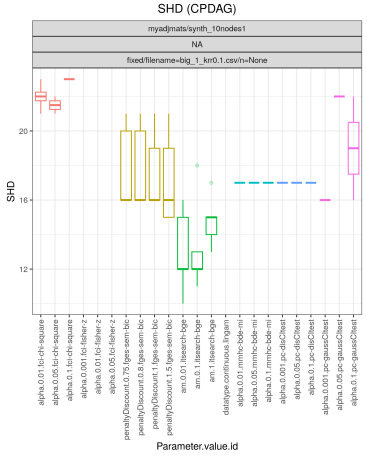
((a)) SHD Scores on the Synthetic 10 nodes data set. Discretized, no noise.



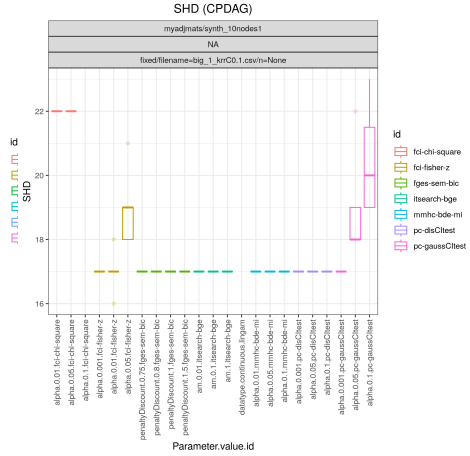
((b)) Synthetic 10 nodes data, Geo C-wise mechanism, max probability 0.1.



((c)) Synthetic 10 nodes data, Geo Comb mechanism, max probability 0.1.

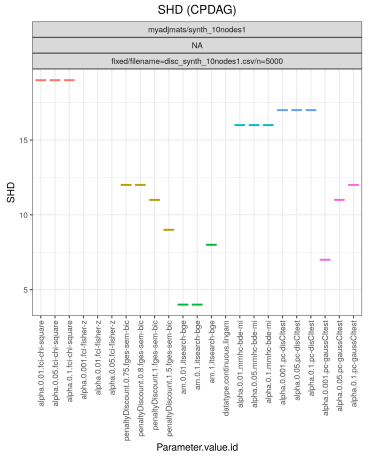


((d)) Synthetic 10 nodes data, k-RR C-wise mechanism, max probability 0.1.

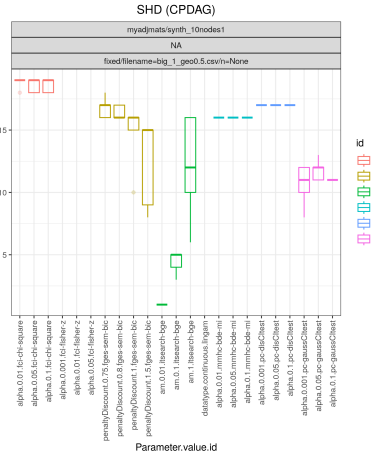


((e)) Synthetic 10 nodes data, k-RR Comb mechanism, max probability 0.1.

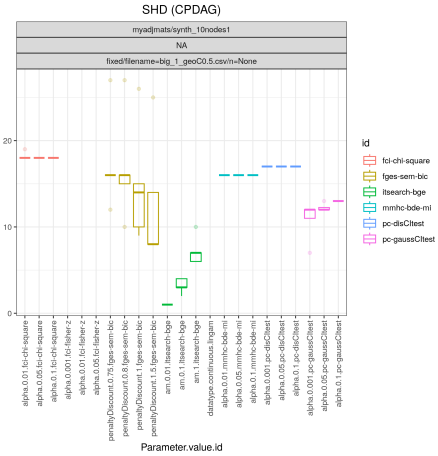
Figure C.29: Synth10 data, SHD, p-max 0.1



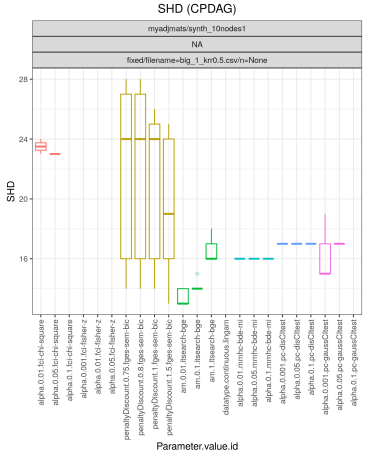
((a)) SHD Scores on the Synthetic 10 nodes data set. Discretized, no noise.



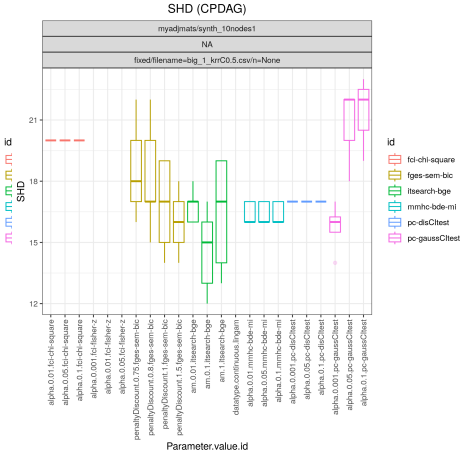
((b)) Synthetic 10 nodes data, Geo C-wise mechanism, max probability 0.5.



((c)) Synthetic 10 nodes data, Geo Comb mechanism, max probability 0.5.



((d)) Synthetic 10 nodes data, k-RR C-wise mechanism, max probability 0.5.



((e)) Synthetic 10 nodes data, k-RR Comb mechanism, max probability 0.5.

Figure C.30: Synth10 data, SHD, p-max 0.5

ECOLE DOCTORALE

Titre : Faire progresser l'IA éthique. Méthodes d'amélioration de l'équité exploitant la causalité et sous contrainte de confidentialité.

Mots clés : IA éthique, équité, confidentialité, explicabilité, causalité

Résumé : L'intelligence artificielle (IA) éthique englobe les pratiques et théories dans le domaine de l'IA et de l'apprentissage automatique (ML) conçues pour s'aligner sur des valeurs morales, abordant leur impact potentiel sur la vie humaine. Les principes éthiques communs de l'IA incluent l'équité, la confidentialité, l'interprétabilité et la fiabilité. Cette thèse explore l'équité, la causalité et la confidentialité dans l'apprentissage automatique et l'IA, considérant la causalité comme un outil pour atténuer les compromis et favoriser les synergies dans l'IA éthique. La contribution de la thèse réside dans l'intégration de méthodes pour l'équité, la confidentialité et la causalité. Des chapitres spécifiques se concentrent sur l'amélioration de l'équité, l'analyse des biais liés à la sous-représentation, la comparaison de mécanismes privés dans la découverte causale, et mettent en avant la nécessité de la causalité dans une IA équitable. La thèse se conclut par une analyse des biais liés au genre et au sexe dans les données de la COVID-19 à travers le prisme de la causalité. Dans l'ensemble, elle apporte des perspectives novatrices pour obtenir de meilleurs compromis entre l'équité, la précision, la confidentialité et l'explicabilité dans les systèmes d'IA, proposant des orientations pour la recherche future et soutenant une approche holistique du développement de l'IA éthique.

Title : Advancing Ethical AI. Methods for fairness enhancement leveraging on causality and under privacy constraint

Keywords : Ethical AI, Fairness, Privacy, Explainability, Causality

Abstract : Ethical Artificial Intelligence (AI) encompasses practices and theories in AI and machine learning (ML) designed to align with moral values, addressing their potential impact on human lives. Common AI ethical principles include fairness, privacy, interpretability, and reliability. This thesis delves into fairness, causality, and privacy in machine learning and AI, viewing causality as a tool to soften trade-offs and foster synergies in ethical AI. The contribution of the thesis lies in incorporating methods for fairness, privacy, and causality. Specific chapters focus on enhancing fairness, analyzing underrepresentation biases, comparing private mechanisms in causal discovery, and emphasizing the need for causality in fair AI. The thesis concludes with a analysis of gender and sex bias in COVID-19 data through the lens of causality. Overall, it contributes novel insights for achieving better trade-offs between fairness, accuracy, privacy, and explainability in AI systems, proposing future research directions and endorsing a holistic approach to ethical AI development.

Institut Polytechnique de Paris
91120 Palaiseau, France